PERSONALIZED DATA-TO-TEXT NEURAL GENERATION

3 juillet 2023

Ben KABONGO

Stage - M1 DAC - Sorbonne Université

Introduction

Data-to-text

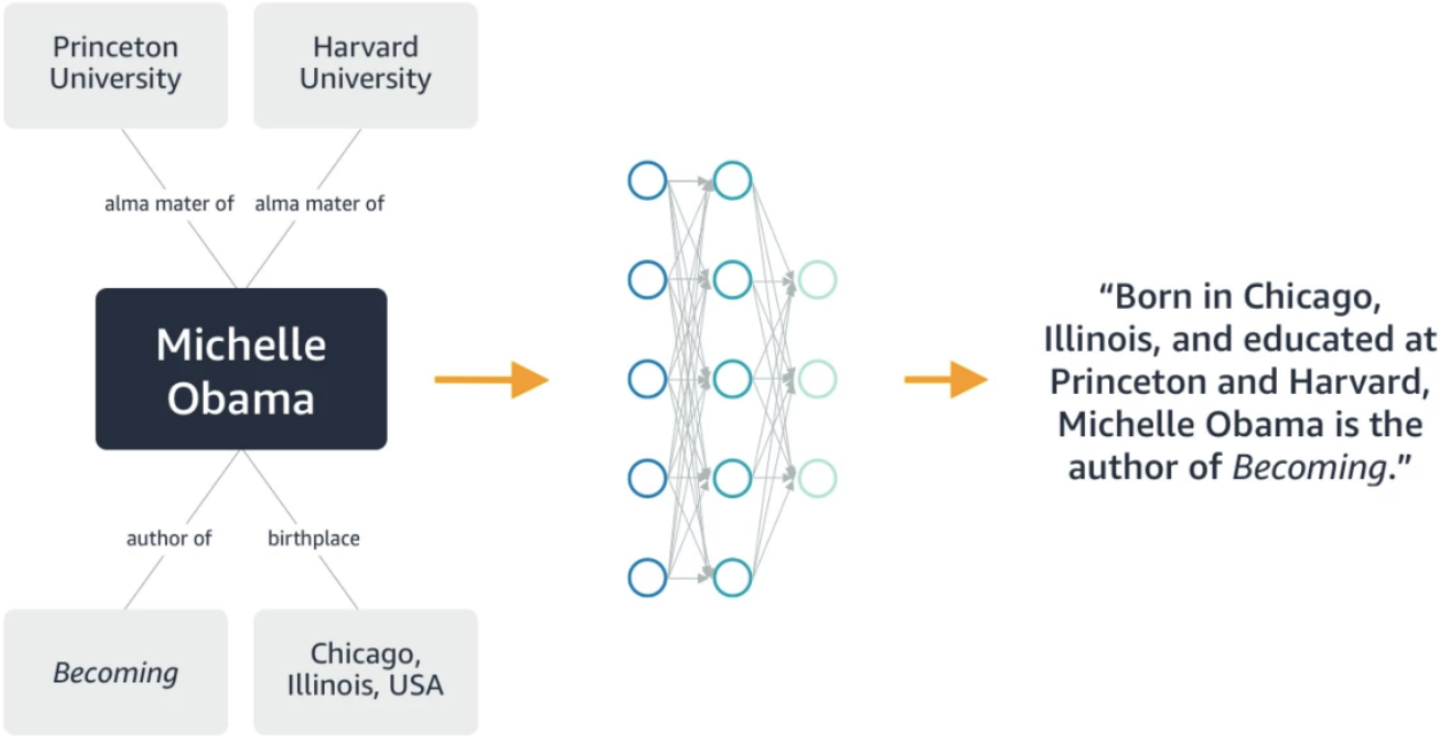


Figure 1 – Data-to-text - illustration

- **Data-to-text** : génération en langage naturel d'une description textuelle pour des données structurées ou semi-structurées (graphes, tables, etc.)

Personnalisation

- **Personnalisation** : tenir compte des préférences de l'utilisateur
- **Style d'écriture** : façon d'écrire d'un utilisateur : expressions, vocabulaire, figures de style, etc.
- **Description personnalisée** : écrite avec le style d'un utilisateur

Data-to-text personnalisé

- **Data-to-text personnalisé** : génération en langage naturel d'une description textuelle **personnalisée** pour des données (semi-)structurées
- **Dataset de data-to-text personnalisé** : $\{(x^i, u^i, y^i)\}_{i=1}^{N}$
  - x^i : données (semi-)structurées
  - u^i : informations sur l'utilisateur
  - y^i : description textuelle personnalisée pour u^i de l'exemple x^i

Problématiques et objectifs

Problématique

- Beaucoup de travaux sur le data-to-text, sur la personnalisation.
- Pas de travaux sur le data-to-text personnalisé
- Il n'existe pas de dataset de data-to-text personnalisé

Objectifs

- **Dataset** : création d'un dataset de data-to-text personnalisé
- **Framework** : proposition d'un modèle de data-to-text-personnalisé

Stage 2022

- Même thématique travaillée l'an dernier. Objectifs compliqués à atteindre.
- **PENS** : dataset et framework de génération de titres d'articles de journaux personnalisé en fonction du corps des articles et des préférences des utilisateurs.
- **Résumé du travail réalisé** :
  - **Création du dataset WikiRoto** : dataset de data-to-text très pertinent pour le data-to-text personnalisé
  - **Data-to-text** : utilisation du modèle T5 pour du data-to-text sur WikiRoto
  - **Personnalisation** : prompt tuning sur les données de PENS Ao et al. 2021

Datasets

Rotten Tomatoes movies and critic reviews dataset

- **Deux datasets** : films et critiques
- **Informations sur des films** : identifiant du film, titre, auteurs, acteurs, directeurs, genres, audience, etc.
- **Critiques utilisateurs** : nom de l'utilisateur, identifiant du film, note (différentes notations et barèmes), contenu, date, etc.
- **1130017** critiques pour **17712** films et **11108** utilisateurs
- **3277** utilisateurs avec au moins 30 avis

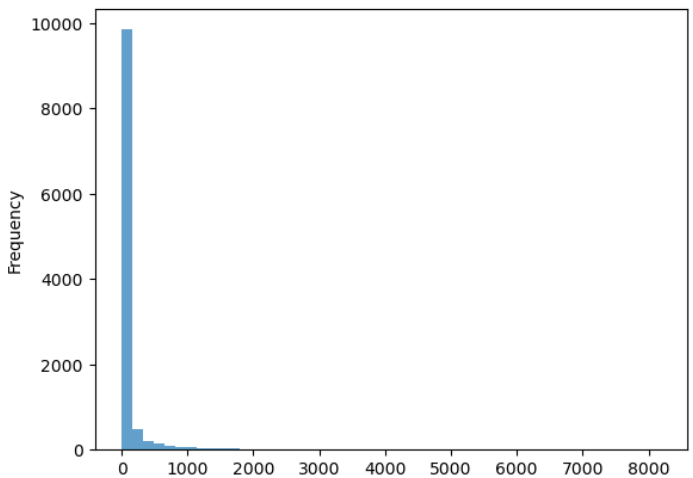


Figure 2 – Distribution des critiques par utilisateur

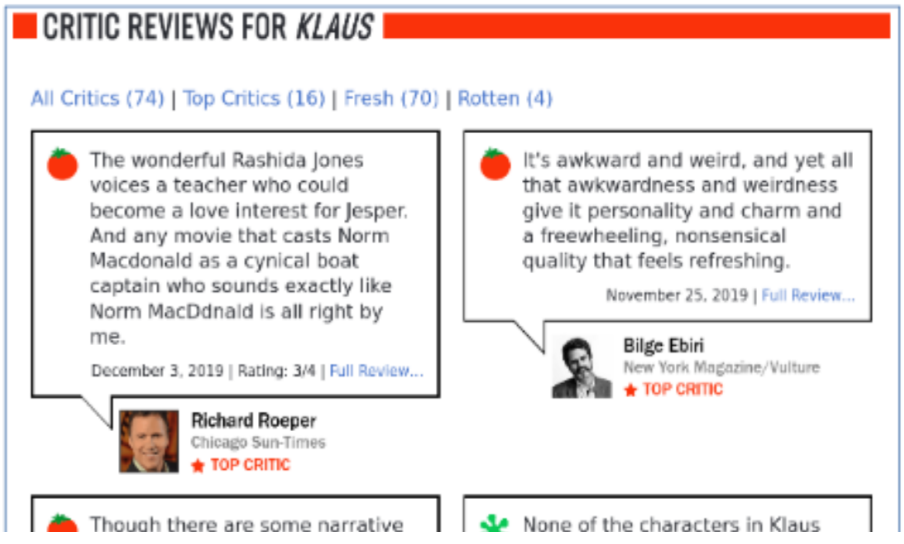
Dataset WikiRoto

Klaus

Release poster

Directed by	Sergio Pablos
Screenplay by	Sergio Pablos Jim Mahoney Zach Lewis
Story by	Sergio Pablos
Produced by	Jinko Gotoh Sergio Pablos Marisa Roman Matt Teevan Mercedes Gamero Mikel Lejarza Gustavo Ferrada
Starring	Jason Schwartzman J. K. Simmons Rashida Jones

Klaus is a 2019 Spanish-American animated Christmas film written and directed by Sergio Pablos in his directorial debut,[2] produced by his company Sergio Pablos Animation Studios and distributed by Netflix. Co-written by Zach Lewis and Jim Mahoney, and co-directed by Carlos Martinez Lopez, the traditionally animated film stars the voices of Jason Schwartzman, J. K. Simmons, Rashida Jones, Will Sasso, Neda Margrethe Labba, Sergio Pablos, Norm Macdonald (in his final film role released in his lifetime), and Joan Cusack. Serving as an alternate origin story of Santa Claus independent from the historical Saint Nicholas of Myra and using a fictional 19th-century setting, the plot revolves around a postman stationed in an island town to the Far North who befriends a reclusive toymaker (Klaus).



CRITIC REVIEWS FOR *KLAUS*

All Critics (74) | Top Critics (16) | Fresh (70) | Rotten (4)

The wonderful Rashida Jones voices a teacher who could become a love interest for Jesper. And any movie that casts Norm Macdonald as a cynical boat captain who sounds exactly like Norm MacDdnald is all right by me.

December 3, 2019 | Rating: 3/4 | Full Review...

Richard Roeper
Chicago Sun-Times
★ TOP CRITIC

It's awkward and weird, and yet all that awkwardness and weirdness give it personality and charm and a freewheeling, nonsensical quality that feels refreshing.

November 25, 2019 | Full Review...

Bilge Ebiri
New York Magazine/Vulture
★ TOP CRITIC

Though there are some narrative

None of the characters in Klaus

Dataset WikiRoto

- Rotten Tomatoes n'est pas un dataset de data-to-text. Pas de description textuelle des films
- **WikiRoto** : Wikipedia + Rotten Tomatoes
- **Data** : informations sur les films
- **Text** : premier paragraphe de texte de la page Wikipedia du film

Méthodologie

Création d'un dataset de data-to-text personnalisé

Problématique

- **Première problématique** : création d'un dataset de data-to-text personnalisé
- Recours à des annotateurs humains : coûteux
- Comment créer **automatiquement** un dataset acceptable ?

Solution

- **Entrée** : **Dataset de data-to-text** : *WikiRoto* + **Dataset de styles d'utilisateurs** : *Rotten Tomatoes*
- **Sortie** : Dataset de data-to-text personnalisé
- **Idée** : dériver pour chaque exemple non personnalisé (x^i, y^i) un exemple personnalisé $(x'i, u, y_u^i)$ pour chaque utilisateur u

Création d'un dataset de data-to-text personnalisé

SCIENCES SORBONNE UNIVERSITÉ

Problématique

- **Modèle de transfert de style** : apprentissage supervisé ou non superivisé ?
- On veut rester proche de WikiRoto
- Impossible de faire du supervisé avec Rotten Tomatoes

Solution

- Transfert de style par apprentissage non supervisé
- **STRAP** : méthode présentée dans le papier de Krishna, Wieting et Iyyer 2020
- Transfert de style = Reformulation = Génération de paraphrase avec **GPT-2**
- **Métriques** : transfer accuracy, semantic similarity, fluency, métrique agrégée, évaluation humaine

Data-to-text personnalisé : grandes étapes

- **Création du dataset**
  - Entraîner et évaluer le modèle STRAP avec les données des styles utilisateur de Rotten Tomatoes
  - Inférer et évaluer le modèle sur les données de WikiRoto
- **Proposition d'un framework**
  - Prompt tuning avec T5 ?
  - Proposition d'autres modèles
  - Etudes du profil utilisateur, de la personnalisation
  - Analyses et évaluation des résultats

Expérimentations et études

Recommandation et analyse des sentiments

- **Recommandation**
  - Recommandation des films aux utilisateurs, en fonction des notes données dans les critiques
  - *Utilité* : comparaison recommandation et sorties de nos futurs modèles
- **Analyse des sentiments**
  - Analyse des sentiments des critiques des utilisateurs
  - **Utilité** : comparaison analyse de sentiments critiques réelles et sorties de nos futurs modèles.

Autres études en cours

- Etudes des profils utilisateurs
- Génération de critique étant donné un film et un utilisateur
- Génération de notes de film étant donné un film et un utilisateur
- Création du dataset