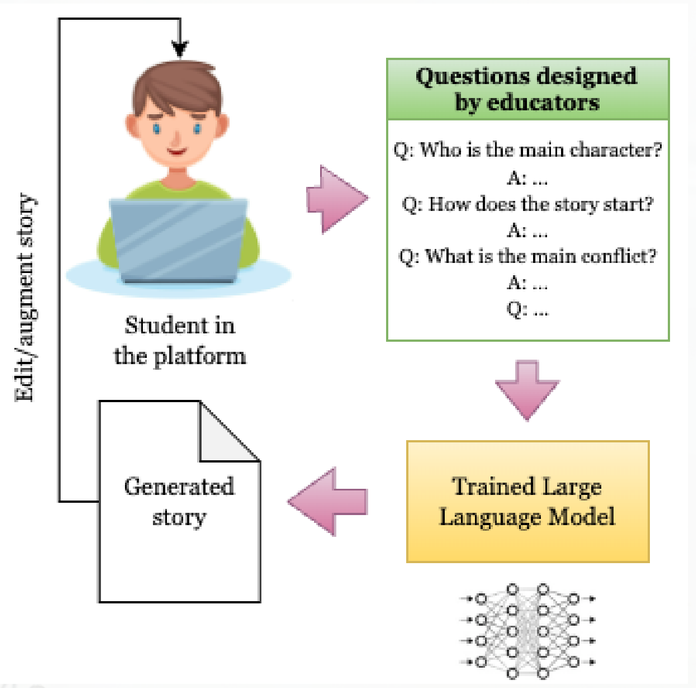
Enhancing Writing Skills of Chilean Adolescents: Assisted Story Creation with LLMs

Hernan Lira*, Luis Martí, Nayat Sanchez-Pi
hernan.lira@inria.cl

Abstract

This study presents an innovative automatic story generation model in Chilean Spanish with the primary objective of enhancing students' writing skills. Writing proficiency holds significant importance in the realm of education, often constrained by limited resources and students' struggles, such as "writer's block." To address these challenges, we propose a machine-learning tool that streamlines story creation, encouraging students to craft narratives aligned with their interests. This is achieved by employing a thoughtfully designed set of questions developed by educators, accompanied by corresponding responses from students, to generate natural language stories. Subsequently, students can iteratively refine and expand their narratives based on this initial framework.



Methods

Data Collection and Corpus Creation

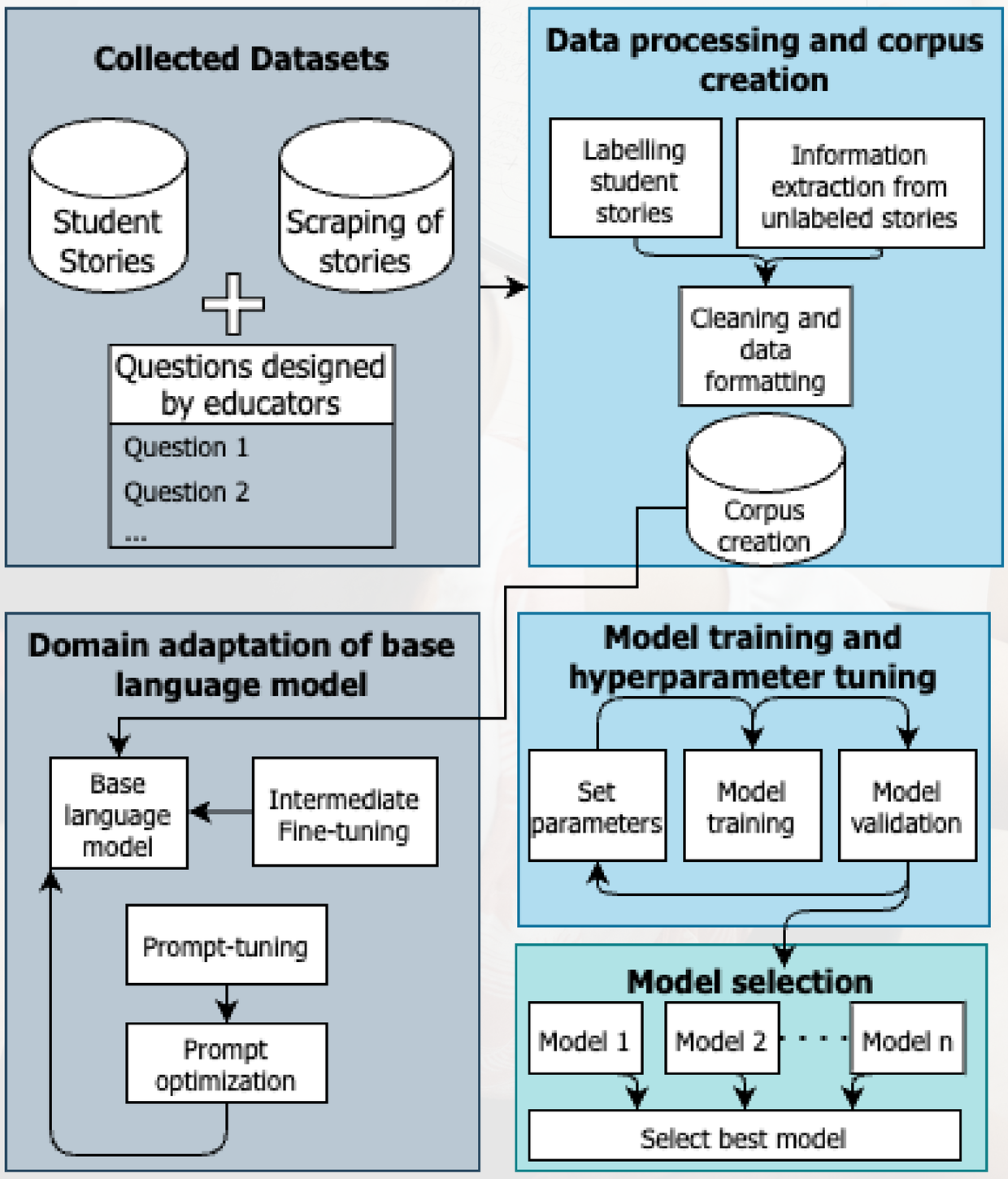
We sourced 769 narratives originating from previous Chilean students assignments. Additionally, web scraping techniques were employed to expand our collection of Spanish narratives.

We establish **associations between the collected narratives and questions** using a method inspired by Veyseh et al. (2021b), employing the "student-teacher" neural network architecture. The "teacher" network learns from labeled narratives to establish foundational knowledge, while the "student" network is trained using a combination of labeled and unlabeled narratives, guided by constraints to ensure label consistency. The corpus ultimately consists of 30,151 stories.

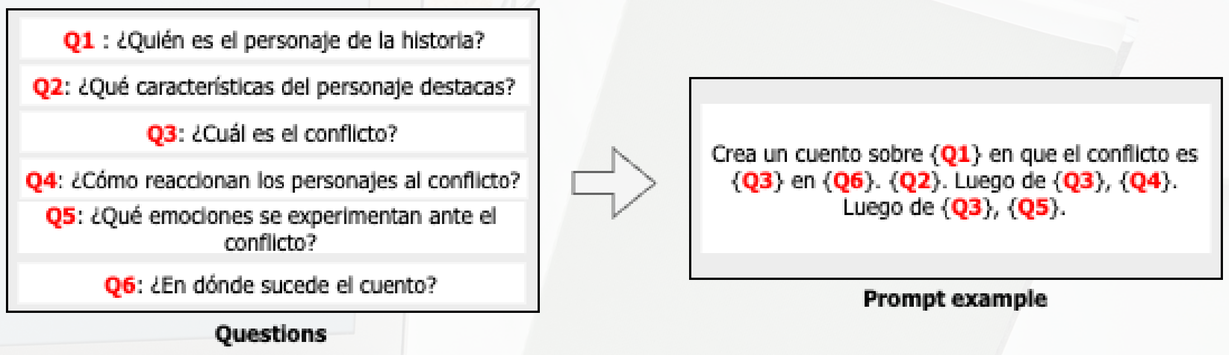
Domain adaptation and training

We employ intermediate fine-tuning for its flexibility, as it allows the corpus to incorporate examples without strict labels or within slightly different domains, thereby expanding the corpus size.

In addition, we use the template-based approach for prompt construction. We draw inspiration from Zhao et al. (2021), who highlight the influence of prompt variation on the model's performance. We generate multiple prompt designs based on templates to assess their effectiveness.



The resulting models are based on GPT-2, GPT-3, and BLOOM, offering multiple versions with varying parameter counts. Domain-adapted base language models are trained on the corpus and optimized through Bayesian optimization. Model validation and selection aims to select the best-performing combination.



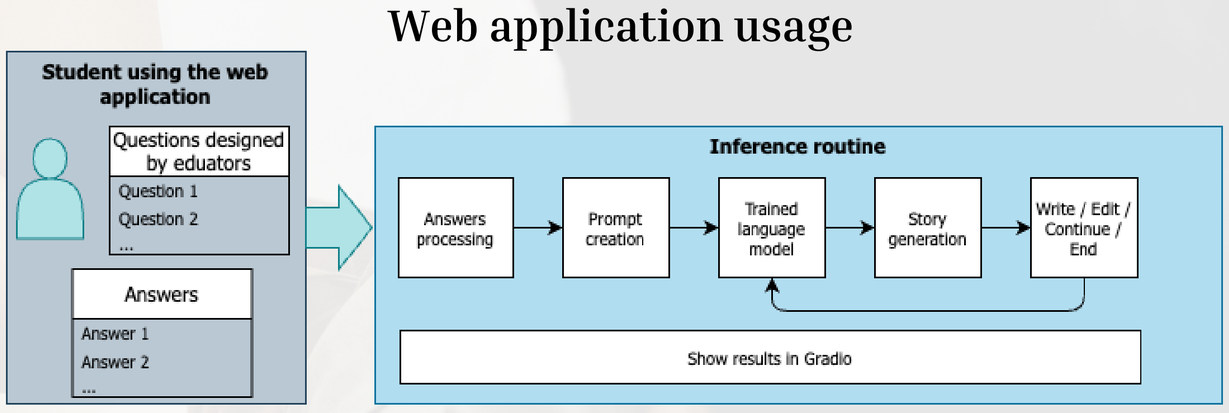
Results

The following are the results of the base models. It's worth noting that GPT-2 and GPT-J have been optimized for the Spanish language. We evaluate their performance using both the entire corpus and the subset containing labeled stories.

	Model	Fluency/Coherence				Divergence
		CBL ⇑	MSJ ⇑	BES ⇑	BRT ⇓	BBL ⇑
Corpus	GPT3	**39.6**	**15.0**	**87.0**	**8.6**	**25.6**
	BLOOM	33.5	14.6	86.8	9.5	24.8
	BLOOM-7b1	26.6	14.6	86.7	9.7	25.0
	GPT-2 Spanish	27.2	11.6	86.6	8.6	24.6
	BERTIN GPT-J	27.5	14.7	86.8	9.5	25.1
Etiqueta	GPT3	**39.3**	**14.9**	**87.1**	**8.8**	**25.3**
	BLOOM	32.3	14.5	86.1	9.8	24.2
	BLOOM-7b1	26.2	14.1	86.0	9.9	25.5
	GPT-2 Spanish	26.8	11.3	85.7	9.0	23.9
	BERTIN GPT-J	26.9	14.2	86.1	9.7	25.0

The GPT-3-based model outperforms others in all metrics. All models show reduced performance when considering only labeled stories, highlighting the need for expanding story collection and information extraction. Notably, GPT-3-generated text exhibits exceptional qualitative robustness, seamlessly integrating responses to prompts.

Web application usage



Human Evaluation

Educators assessed our model with input from 10 Chilean students and adults. They evaluated coherence, fluency, grammar, relevance, and human-likeness. **The model excelled in coherence, fluency, grammar, and relevance, meeting linguistic and storytelling requirements effectively.** Human-likeness, a subjective metric, revealed potential areas for improvement.