Simplifying College Admission Texts

CS5803 - NATURAL LANGUAGE PROCESSING

Group-7

Chintalapudi Abhiroop – AI20BTECH11005

Nelakuditi Rahul Naga – AI20BTECH11029

Vaddamani Saketh – CS20BTECH11054

Introduction

- Access to the higher education system for minoritized populations, especially low-income students, first-generation in college students, and students whose first spoken language is not English, is an important social challenge.
- Researchers have consistently explained that much of higher education's communication with prospective students is too complex, too lengthy, and requires a wealth of prior knowledge of the higher education system to successfully navigate.
- Many institutions publish admissions application instructions far above the average reading level of a typical high school graduate, making it hard for the average prospective student to read and comprehend the instructions.
- The verbose, difficult communication places an unnecessary barrier between students and access to higher education.
- We can utilize text simplification models to tackle this challenge.

Problem Statement

- There have not been much studies regarding how to simplify admissions process and make them more accessible for prospective students without losing important details.
- Most of the existing work in text simplification largely relies on established datasets in the News and Wikipedia domains.
- Unlike the News and Wikipedia domains, the jargon and concepts encapsulated within admission texts are not ones that a person encounters every day. Hence these texts are more specialized and difficult to generalize.
- The performance of text simplification models in this domain is not great because of the lack of diversity in datasets in this domain.
- Our goal is to compare the performance of various baseline models (both non-finetuned and finetuned) to the SimSum model proposed in Sofia et al. on College admission specific documents.

Dataset

- To train and evaluate our models, we utilized **PSAT** (**P**rofessionally **S**implified **A**dmissions **T**exts), which was curated from scratch by Taylor et al.
- It is the first manually simplified and verified corpus of admissions instructions sourced from 112 randomly sampled US post-secondary institutions.
- Every document was manually verified by 2 subject-matter experts (among a total of 10 experts) who are employed as full-time admissions professionals in US institutions of higher education.
- The simplified documents reduced the reading level of these texts from grade 13.3 to grade 9.8, making it much more accessible for minority and emergent bilingual students.

(Original) You must complete the following steps before USF will consider your application complete and begin admission evaluation. Pay the non-refundable $30 application fee or submit an application fee waiver.

(Simplified) You must submit an online application with a nonrefundable $30 application fee. You can also submit an application fee waiver.

Data Pre-processing

- For each of the 112 universities in the PSAT dataset, there is a separate original and simplified document.
- We first split both the original and simplified document data into 50% train, 30% test, and 20% validation datasets using pre-defined indexes provided by the authors of PSAT paper.
- This gave 56, 33 and 23 original and simplified documents for train, test and validation datasets, respectively.
- Each of the train, test and validation datasets from original and simplified sets were then converted into **.complex** and **.simple** files, respectively. Thus, a total of 6 files were created.
- This was achieved through concatenation of various documents (after stripping new lines in each document) line by line.

Models and Implementation Details

- The models that we used for text simplification are:
  - BART (finetuned and un-finetuned) - 139M parameters
  - T5 (finetuned and un-finetuned) - 222M parameters
  - BRIO - 406M parameters
  - SimSum Model with BART backbone - 278M parameters
- The base models for BART, BRIO and T5 were imported from **HuggingFace** using **transformers** library. The code for finetuning these models was implemented using **PyTorch Lightning** library.
- SimSum BART model was also implemented using PyTorch Lightning.
- The finetuning for all the above models was performed for a duration of 3 epochs using a Nvidia Tesla T4 GPU provided by Google Colab.

Evaluation Metrics

- **SARI:** This metric compares the system output against references and against the input sentence, which explicitly measures the goodness of words that are added, deleted, and kept by the systems.
- **D-SARI:** This metric is a modified SARI score with additional penalty factors based on text length and specially designed for the document-level text simplification task.
- **BLEU:** This metric is commonly used in machine translation and other conditional generation tasks. It has a strong correlation with grammar and meaning.
- **FKGL:** This metric is used to measure readability but does not consider grammar or meaning preservation.
- SARI, BLEU and FKGL metrics are computed using EASSE, a Python3 package created to standardize automatic evaluation and comparison of sentence simplification systems. D-SARI is computed using a customized function.

Performance Comparison

Model	Metrics			
	SARI↑	D-SARI↑	BLEU↑	FKGL↓
Untuned BART	31.96	19.58	0.2239	10.15
Untuned T5	32.12	28.45	0.0304	7.17
BRIO	45.25	26.93	**0.2258**	6.41
BART	40.22	19.20	0.1356	6.22
T5	43.23	24.84	0.1556	8.06
SIMSUM BART	**47.56**	**36.45**	0.1894	**5.80**

- We can observe that SIMSUM model with BART as backbone, fine-tuned on the PSAT dataset for 3 epochs, achieved best results on SARI, D-SARI and FKGL metrics.

Sample Results

- **Original (Complex):** You may apply for admission to Missouri S&T any time after completing six semesters of high school coursework (typically after your junior year). When you submit your application for admission, you are also applying for merit-based university scholarships. To complete the application process, be sure to complete the steps detailed below. Students can be admitted on unofficial transcripts and test scores. 1.) Complete your online application It is important to remember the email address and password you used to create your application account. 2.) Missouri S&T does not charge an application fee! 3.) Submit your high school and/or college transcripts Transcripts can be sent directly from your high school and/or college to Missouri S&T. MAIL TO : Missouri University of Science and Technology 106 Parker Hall, 300 W 13th Street Rolla, Missouri USA 65409-1060 4.) Submit your ACT/SAT test scores Test scores can be sent directly from the testing agency. Use these codes to send test scores to S&T: ACT: 2398 || SAT: 6876
- **Reference (Simple):** How to Apply: You must submit an online application. The application is free. Required Materials: High school transcript(s). College transcript(s) (if any). You can submit transcripts from your high school and/or college to: Missouri University of Science and Technology 106 Parker Hall, 300 W 13th Street Rolla, Missouri USA 65409-1060 Official ACT and/or SAT scores. Our codes are: ACT = 2398 SAT = 6876
- **SimSum BART Output:** How to Apply: You must submit an application using the Missouri S&T Application. Required Materials: Official high school and/or college transcript(s). If you have questions, you can contact us at: Missouri University of Science and Technology 106 Parker Hall, 300 W 13th Street Rolla, Missouri USA 65409
- **Color Codes:** Green - Model covered these parts, Red - Model ignored these parts, Blue - Model replaced these parts

Conclusions

- Accessing higher education systems is an important social challenge for the case of minoritized populations, since most of them required being at least at 14th grade English reading comprehension level.
- Modern conditional text generation models that perform automatic text simplification are a plausible way to make higher education more accessible to the underprivileged populations.
- In our work, we compared the performance of various text simplification models like BART, T5, BRIO, SIMSUM on the PSAT dataset, the first manually simplified and verified corpus of college admissions.
- SIMSUM model with BART as backbone achieved the best results on SARI, D-SARI and FKGL metrics.

Limitations and Future Works

- There are over 6,000 institutions of higher education in the US, PSAT dataset sampled only a small number of these institutions.
- The authors of PSAT paper said that the size of PSAT was confined mainly due to time constraints and the huge amount of work involved to gather, simplify text and work with subject-matter experts to approve simplified texts.
- In future, we can work with bigger datasets, sampled from more academic institutions around the world, not only US. This will ensure that the text simplification models have better generalization capability.
- We can also try to use various other kinds of models and further improve text simplification in this domain.

References

- SIMSUM: Document-level Text Simplification via Simultaneous Summarization, Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian, ACL 2023.
- Text Simplification of College Admissions Instructions: A Professionally Simplified and Verified Corpus, Zachary Taylor, Junyi Jessy Li, Maximus Chu, 2022.
- SimSum Code - GitHub


THANK YOU