Simplifying College Admission Texts

Abhiroop Chintalapudi
Artificial Intelligence
IIT Hyderabad
AI20BTECH11005

Nelakuditi Rahul Naga
Artificial Intelligence
IIT Hyderabad
AI20BTECH11029

Vaddamani Saketh
Computer Science
IIT Hyderabad
CS20BTECH11054

Abstract—**For students whose first language is not English and are members of minority populations, getting into higher education has become a significant societal barrier. This is a result of higher education institutions providing prospective students with instructions that are frequently overly complicated, drawn out, and require a great deal of prior knowledge to adequately navigate.**

There has been less research on how to simplify and make college admissions instructions easier to understand for potential students without sacrificing critical information. The majority of the current research in text simplification is mostly dependent on well-established datasets from the Wikipedia and News domains. In contrast to the News and Wikipedia domains, admission texts contain topics and vocabulary that are not commonly seen by readers. These texts are therefore more specialised and challenging to generalize. Because there is little variability in the datasets within this domain, text simplification models struggle to perform well. In our paper, we compare the performance of various baseline models (both non-finetuned and finetuned) to the SimSum model proposed in Sofia et al. [1] on PSAT (Professionally Simplified Admissions Texts) [2] , a dataset created from admission instructions of 112 randomly selected higher education institutions in US. PSAT was curated, verified and accepted by experts who work in admission offices at various US institutions. We demonstrate that SimSum model (with BART backbone) outperforms other existing baseline models on this dataset.

I. Introduction

For minoritized populations, including low-income students, first-generation college students, students of colour, and students whose first language is not English, access to the higher education system is critical. Research in this area, however, revealed that the communication between these prospective students and higher education institutions is overly complex. Additionally, the instructions published by these institutions for submitting admission applications read at or above the English reading comprehension level of the 14$^{\text{th}}$ grade, which is far too high for the average adult or prospective student attempting to pursue higher education. This communication gap and the difficult verbiage employed to express ideas are putting an unneeded obstacle in the way of students' admission to higher education.

With the recent explosion of advances in the field of natural language processing, text simplification is a plausible way to tackle this obstacle. Text simplification attempts to retain most of the information in the original text while making a document's content easier to understand and accessible to all readers, regardless of background or reading level. Text simplification employs less complex syntax and sentence structures, as well as a simplified vocabulary with definitions and explanations for any difficult terminology used. It is regarded as a sequence-to-sequence text generation problem, and is similar to other NLP generation tasks like paraphrasing and text summarising.

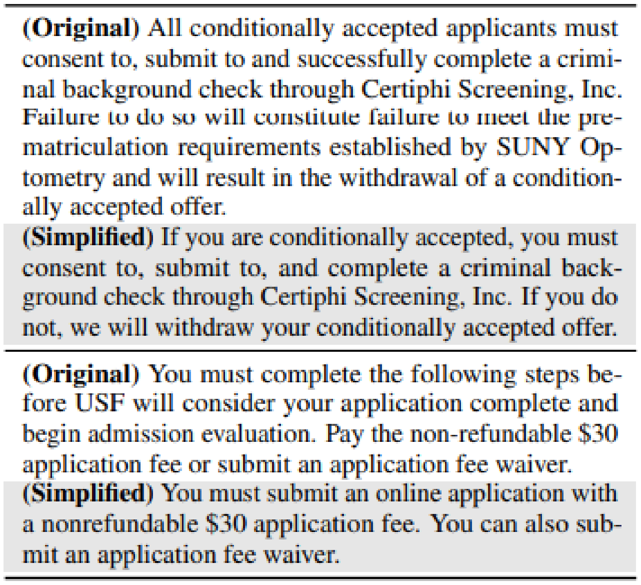
We discovered that there aren't many studies out there right now exploring ways to make college admissions guidelines simpler while keeping majority of the crucial information in order to make the application process easier for potential students and their families. It was also discovered that a significant portion of the current research on text simplification heavily relies on well-established datasets from the Wikipedia and News domains. However, in contrast to these two areas, college admission texts are thought to be more specialised and challenging to generalize due to their unique terminology and concepts used, which we don't encounter everyday.

As a result, we don't know much about the nature of text simplification in the context of college admissions or how different text simplification models perform in this new field. Therefore, we compare the performances of several baseline models (with and without fine-tuning) to the recently developed SimSum model in the domain of college admission texts.

II. Dataset

A. Professionally Simplified Admission Texts (PSAT)

In our studies, we compared the text simplification models' performances on the PSAT (Professionally Simplified Admissions Texts) dataset, which was specifically designed for the domain of simplifying college admission texts. PSAT is the first manually simplified and validated corpus of admissions instructions from 112 randomly selected post-secondary educational institutions in the United States. The substantial existing literature of research on manual simplification - which includes reducing syntactic complexity, enhancing lexical cohesiveness, and expounding on jargon - served as the basis for this professional simplification method. Two subject-matter experts who work full-time as admissions officials in US higher education institutions manually verified each document in PSAT. Figure 1 illustrates a sample of original and simplified texts from the PSAT dataset.



(Original) All conditionally accepted applicants must consent to, submit to and successfully complete a criminal background check through Certiphi Screening, Inc. Failure to do so will constitute failure to meet the prematriculation requirements established by SUNY Optometry and will result in the withdrawal of a conditionally accepted offer.

(Simplified) If you are conditionally accepted, you must consent to, submit to, and complete a criminal background check through Certiphi Screening, Inc. If you do not, we will withdraw your conditionally accepted offer.

(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.

Fig. 1: Original vs Expert Simplified Versions of College Admissions Texts

It was found that the simplified documents in PSAT reduced the reading level of the texts from grade 13.3 to grade 9.8, making it much more accessible for minority and emergent bilingual students.

B. Data Preprocessing

We converted the PSAT dataset into appropriate format required for training the text simplification models. The step by step process we followed is as follows:

1) 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.
2) 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.
3) This was achieved through concatenation of various documents (after stripping new lines in each document) line by line.

III. Experiments

A. Models and Implementation Details

The details regarding the models that we used for text simplification are as follows:

1) **BART:** This is an effective sequence to sequence model that achieves excellent results on various text summarization tasks [3]. We evaluate both untuned and finetuned BART models on PSAT dataset. The pre-trained BART model (facebook-bart-base) is imported from HuggingFace using transformers library. The model has 139M trainable parameters.
2) **BRIO:** This is also a pre-trained model with top performance on various sequence-to-sequence tasks [4]. The pre-trained BRIO model (Yale-LILY/brio-cnndm-uncased) is imported from HuggingFace using transformers library. We fine-tune this model on the PSAT dataset. The model has 406M trainable parameters.
3) **T5:** This is an encoder-decoder model proposed by Google [5]. It is pre-trained on a multi-task mixture of unsupervised and supervised tasks. We evaluate both untuned and finetuned T5 models on PSAT dataset. The pre-trained T5 model (google-t5-base) is imported from HuggingFace using transformers library. The model has 222M trainable parameters.
4) **SIMSUM BART:** In this model, we use the pre-trained version of BART (facebook-bart-base) as the backbone of SimSum model on simultaneous summarization and simplification stages. The model is finetuned on PSAT dataset. The model has 278M trainable parameters.
5) **SIMSUM T5:** In this model, we use the pre-trained version of T5 (google-t5-base) as the backbone of SimSum model on simultaneous summarization and simplification stages. Two versions of this model are finetuned on PSAT dataset - Vanilla SimSum T5 and SimSum T5 model with Embedding Similarity loss. The model has 445M trainable parameters.

The code for finetuning above models was implemented using **PyTorch Lightning** library. SimSum BART and T5 models were also implemented using PyTorch Lightning. The reference code for implementing the models was taken from [6], it was developed by the Authors of [1].

Finetuning for all the above models was performed for a duration of 3 epochs using a Nvidia Tesla T4 GPU provided by Google Colab. All other hyperparameters were kept the same as in [1] while training the models.

B. Evaluation Metrics

We have used the following standard metrics for evaluating the performance of our text simplification models:

1) **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.
2) **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.

3) **BLEU:** This metric is commonly used in machine translation and other conditional generation tasks. It has a strong correlation with grammar and meaning.
4) **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.

C. Performance Comparison

The evaluation metrics for various text simplification models on PSAT are shown in table I below:

Model	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 T5 (Vanilla)	33.31	21.89	0.0613	6.47
SIMSUM T5♠	34.53	21.28	0.0653	7.17
SIMSUM BART	**47.56**	**36.45**	0.1894	**5.80**

TABLE I: Evaluation Metrics on PSAT, ♠: SIMSUM T5 model with Embedding Similarity loss

We can observe from above table 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.

D. 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

The color codes are: Green - Model covered these parts, Red - Model ignored these parts, Blue - Model replaced these parts.

IV. 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.

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 intend to 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

[1] Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff and Seyed Ali Bahrainian, “SIMSUM: Document-level Text Simplification via Simultaneous Summarization”, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pages 9927–9944, 2023.

[2] Zachary W. Taylor, Maximus H. Chu, and Junyi Jessi Li, “Text Simplification of College Admissions Instructions: A Professionally Simplified and Verified Corpus”, Proceedings of the 29th International Conference on Computational Linguistics, pages 6505–6515, 2022.

[3] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension“, 2019.

[4] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig, ‘Brio: Bringing order to abstractive summarization“, 2022

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer“, 2019.

[6] SimSum GitHub Code