Text Summarisation and Simplification

Authors (Group 39):
Atin Sakkeer Hussain (A0225782Y), Aayush Mathur (A0218545B), Gupta Ananya Vikas (A0226576W)
Jai Lulla (A0226609B), Tanishq Sharma (A0226569R)

Abstract

Reading research papers is something for which most students struggle for hours, let alone understanding and extracting key information from them. Our project aims to find the best Machine Learning summarisation model that simplifies content-heavy research papers to a concise, yet informative summary for the reader. We specialized state-of-the-art Recurrent Neural Network and Transformer models to our problem by training it with the "Simple English Wikipedia" data set. We then conducted quantitative assessments to measure precision and recall, followed by qualitative data collection via surveys, compared to industrial defined summarising and simplification tools. We found that the Gated Recurrent Unit (GRU) performed the best to solve the problem.

Problem Definition

- Task: To generate simplified summaries of research papers and academic articles
- Target Users: University students
- Resources: Research Papers, Articles, Academic papers
- Proposed Models: RNN, LSTM, GRU, Transformer
- Metrics Analysed: Bleu and rouge scores, accuracy, brevity, and clarity

Motivation

Most of the initial ideas revolved around our daily pain points as university students. We came up with the idea of text summarisation as we all resonated with the pain of scanning through dozens of research papers to find necessary content with ease. In the process, the information is too complex to understand or it is extremely long, increasing the chances of us skipping key sections.

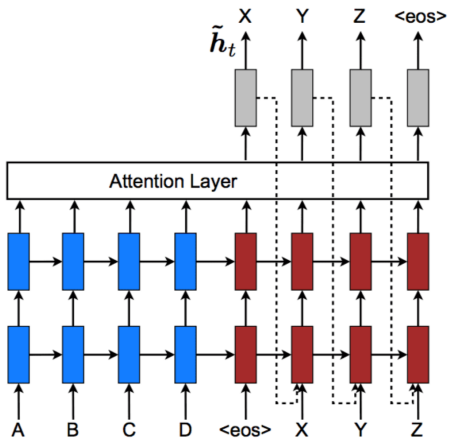
If we look at traditional text summarisers, some of the information gets lost. This could be nomenclature, definitions, and data. This is because they often overlook the simplification aspect - which leads to information loss.
Hence, our focus is to develop a summariser that simplifies texts while addressing these aforementioned issues for students, to optimise the time invested in reviewing each research paper.

Models and Machine Learning Approach used

Considering the nature of the task, we decided to analyze and compare the performance of the following state-of-the-art models:

1) Recurrent Neural Network (RNN) Encoder Decoder with Attention Architecture:

RNN networks are excellent at capturing context information and processing text vectors.



RNN Architecture

Encoder: The RNN Encoder reads the text vector token by token and captures the contextual information of the input vector into a context vector. This is expected to contain good

information regarding the text.
Every RNN cell in this Encoder has a hidden state as well as a cell state. The last hidden and cell state is used as the initializer for the decoder.

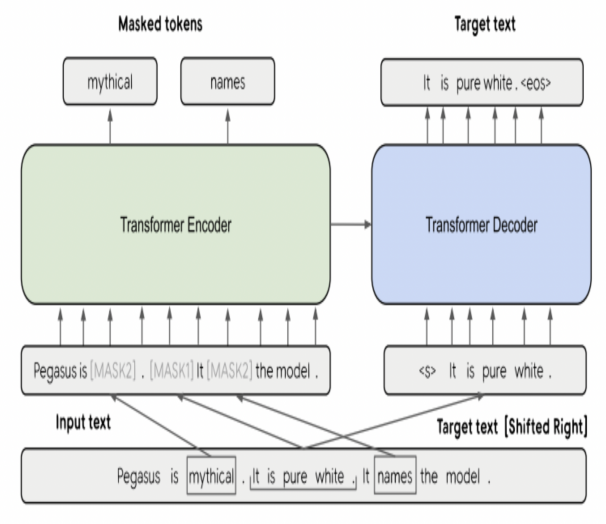
Attention Mechanism: This has the ability to obtain significance in sequences. Using this mechanism, the decoder is able to learn which words to give more 'attention' to when producing the output vector.

Decoder: The RNN Decoder reads the output sequence offset by one along with the last hidden state and cell state of the encoder and predicts the next word in the target sequence. We add [SOS] to the start to indicate start of the sequence and [EOS] to the end to indicate the end of sequence.

We attempted this architecture with 3 different kinds of RNNs as they recognize data sequential characteristics well, which is important for this problem application.

- Simple Recurrent Neural Network (RNN)
- Long Short Term Memory (LSTM)
- Gated Recurrent Unit (GRU)

2) Transformer Architecture:



Transformer Architecture

For this project we chose to use the PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) model, since it has shown to give excellent performance in terms of summarization tasks. This model, similar to other Transformers, consists of an Encoder and Decoder. Additionally, it uses Gap Sentences Generation as a pre-training task to further improve the model's performance.

Experiment Setup

Wikipedia Text Simplification Dataset

Simple English Wikipedia contains articles with similar content to English Wikipedia but with simpler vocabulary and grammar. We utilise the Wikipedia text simplification dataset, which consists of 167K aligned sentence pairs generated by sentence-aligning Simple English Wikipedia and English Wikipedia.

It is a large-scale dataset collated from an online knowledge base written by different authors. The structure of the dataset allows us to train the model on sentences and their simplified equivalents. Since it spans a wide range of topics and styles, we used this dataset to help our model perform better when simplifying the text.

Whether data is representative of our target population?

Since, the aim is simplifying research articles, we chose factual articles to train and test our models to make our approach applicable for students.

In the educational domain, there are data sets available that either contain difficult annotations or have different variants representing different readability levels hence representing our required audience well.

Model parameters, libraries

- Simple Recurrent Neural Network (RNN)
  Library: PyTorch
  Hidden Size: 600
  Epochs: 1000
- Long Short Term Memory (LSTM)
  Library: PyTorch
  Hidden Size: 600
  Epochs: 1000
- Gated Recurrent Unit (GRU)
  Library: PyTorch
  Hidden Size: 600
  Epochs: 1000
- Transformer
  Library: HuggingFace

Hidden Size: A size of 600 was chosen as it gave the best results for simplifying the text, given limited computing power. Exceeding this value caused a memory leak.

The implementation for the models can be found on https://github.com/crypto-code/CS3244-Project

Quantitative Analysis

We defined quantitative success of our proposed model using two state-of-the-art metrics:

- Bleu algorithm to measure precision
- Rouge to measure recall and precision

Model	Bleu Score	Rouge-1 F1-Score	Rouge-2 F1-Score	Rouge-l F1-Score
RNN	0.76	0.92	0.89	0.92
LSTM	0.77	0.93	0.90	0.93
GRU	0.79	0.95	0.90	0.94
Transformer	0.20	0.56	0.29	0.41

Table 1.

Based on Table 1, it is seen that GRU gives the best performance in terms of Bleu and Rouge scores. Although Transformers are state-of-the-art, their poor performance in comparison to RNN models may be attributed to the fact that the dataset used in this project is for simplification rather than just summarization. In a simplification task, the model should also be able to rephrase sentences and, in some cases, even generate new sentences to give the perfect simplification.

Transformers, particularly the PEGASUS model, have difficulty doing this since they focus more on picking out important sentences and words in context from the input sentences to generate the output. RNNs on the other hand do not face this issue since they directly learn an input to output mapping function.

Qualitative Analysis

We conducted a qualitative assessment of our models using a three-step process:

- A double-blind survey conducted internally to compare the accuracy, brevity, and clarity of the summarised texts generated by the four models.
- Choosing the best model using quantitative and qualitative data collected.
- A survey conducted for 17 students to compare the accuracy, brevity, and clarity of text excerpts from the original, best model, and industry-standard model.

  The sampled text extract was taken from our chosen dataset so as to keep consistency of both qualitative and quantitative results. The original and summarized extracts can be seen in the Appendix.

Survey

The survey serves as a seamless way for us to collect qualitative data from potential users. The structure is as follows:

- Original and summarised text(s) (max. 100 words)
- Ranked scale of 1-4 to measure text's:
  (Rank 1: highest , Rank 4: lowest)
  - Accuracy: Ensures the common words with different meanings aren't confused in academic writing and the context of the text is preserved
  - Brevity: Ensures unnecessary information is avoided which is the goal of the model while simplifying and summarising
  - Clarity: Ensures that the context is well-communicated and effectively understood by the users.

Double-blind assessment

An internal survey was conducted amongst 5 team members. The results are as follows:

Score	Accuracy						Brevity						Clarity					
Model/User	1	2	3	4	5	Avg	1	2	3	4	5	Avg	1	2	3	4	5	Avg
RNN	3	4	3	3	3	3.2	3	3	2	2	1	2.2	4	4	4	4	3	3.8
LSTM	1	2	2	2	2	1.8	2	2	3	4	3	2.8	1	3	2	3	2	2.2
GRU	2	1	1	1	1	1.2	1	1	1	1	2	1.2	2	1	1	1	1	1.2
Transformer	4	3	4	4	4	3.8	4	4	4	3	4	3.8	3	2	3	2	4	2.8

Table 2.

Choosing the best model

Table 1 and Table 2 suggests that the GRU model performed the best during quantitative and preliminary qualitative assessment. Hence the model was chosen for further qualitative assessments with our potential users.

User assessment

Further assessment was carried out for 17 users to measure effectiveness of the models compared with online simplification portals and the average ranks are presented in the table below:

Model/Score	Accuracy	Brevity	Clarity
GRU	2.0	2.1	1.3
Simplish.org	2.4	2.4	2.1
QuillBot	1.6	1.5	2.6

Table 3.

Based on user feedback results shown in Table 3, we found GRUs to score better on the clarity metric as compared to industry text simplifiers, "Simplish.org" and "QuillBot." Though the latter are better on accuracy and brevity, our chosen model was better solving the problem of producing a clear summarization for users.

Conclusion

Through our experiments, we have gained valuable insights into creating machine learning models for text simplification. We created two different approaches and four models for the different interpretations of the task. We used the same data for each to compare the performance using quantitative and qualitative metrics and received different results. More training data, computational power, and time for training would have certainly made our results better.

Applications for Singapore

Looking at the big picture, apart from being a great tool for students, there are many industries where our model approach can be a valuable asset.

- *Healthcare Industry:* The healthcare industry is the one of the highest in information generating industry as patients are given instructions on their regimen and at such points, patients either face the problem of having excess information or the instructions are too complex to understand as too many medical terms are included.

  Since Singapore has an increasing aged population, this problem of information management is widely seen in Singapore. Singapore's aged population has been estimated to grow to 25 percent by 2030, and according to a study in 2017, only 57 percent of patients truly understand the instructions disseminated, hence, our model approach can be a direction to simplify and summarise the instructions for the patients and enhance the patient experience for Singapore residents.

- *Law firms:* The core value of our model which is simplification is one of the most relatable necessities for the law firms. Since laymen cannot understand the legal terms and their consequences, Singapore citizens face barriers of communication in such scenarios.

  Hence, we would want to tweak and apply our model approach for Singapore citizens to understand the legal terms and bridge the gap between understanding and lack of knowledge. Simplifying such legal documents like academic documents, will ease the pain of Singaporeans when they have to deal with such serious issues, given the lack of time.

Future Works / Improvements

The future scope of our project would be to further train the models we have already created for more iterations and with greater variety of data, preferably extracted directly from research papers. We would want to try various kinds of texts for simplification to compare the performance. Additionally, we would like to implement more deep learning technologies for exploring the scope of our models better and enhancing its ability to simplify research papers.

Reflection

While brainstorming ideas for the project, we had different ideas with regard to the simplification of the text, and where it could be applied. Looking at the existing applications that are present, they did not simplify to the level, where the text became more understandable. In addition, there were many model approaches which would summarise the texts however, the simplification factor was missing and vice versa. Hence, we wanted to bring the collaboration of both aspects in our solution.

Furthermore, we were also struggling to figure which metrics would add most value to our approach given that textual data can be difficult to classify and summarise as the sequence and context of words needs to be preserved. Hence, after reading several articles and with the feedback of our professor, we chose accuracy and precision as our core metric for determination. We were also ambiguous on which machine learning model would be best suited, as every literature had a varying idea to tackle the problem. Finally, we settled upon the approach of neural networks as it would preserve the sequence of text data as well as give valuable output which can be compared with online portals and industry standard summarising and simplifying tools.

Within our group, each of us implemented one of the proposed models, which involved pre-processing, training, fine-tuning, and summary generation. We spent time discussing challenges and improvements for each of our models, forcing us towards deep self-learning the process. We all played a part in finding users to test our solutions and using data to deduce the best model to tackle this problem. With 4 Computing Students and 1 data science student, we had a lot of expertise and different perspectives into the

problem approaches, which helped us explore our models better along with having factual and structural coherence to each approach.

All in all, this project has opened our eyes to the practical decision making processes of a Machine Learning project. Such learnings can only be done by deep-diving into challenges and using a step-by-step approach to justify proposed solutions. It further helped us question every step of our approach in terms of the pros and cons of implementing a certain method.

Literature Review

We referred to some informative articles to guide our approach and justify our motivation:

Extracting highlights of scientific articles: A supervised summarization approach - Cagliero, L, La Quatra, M.

A Long Texts Summarization Approach to Scientific Articles - Cinthia M. Souza, Renato Vimieiro

A Survey on Text Simplification - Punardeep Sikka, Vijay Mago Semantic Structural Evaluation for Text Simplification - Elior Sulem, Omri Abend, Ari Rappoport

Text Simplification Tools: Using Machine Learning to Discover Features that Identify Difficult Text - D. Kauchak, O. Mouradi, C. Pentoney and G. Leroy

I. Abdul Halim Zaki, R. M. Razali, S. Gnanasan, R. Alias, and M. Karuppannan, "Medication discrepancies among elderly patients discharged from a tertiary hospital: Prevalence and risk factors," Singapore Medical Journal, pp. 362–365, 2021.

D. R. Calkins, "Patient-physician communication at hospital discharge and patients' understanding of the postdischarge treatment plan," Archives of Internal Medicine, vol. 157, no. 9, pp. 1026–1030, 1997.

Appendix

Samples Texts and the outputs for Assessment:

- Original:
  Hox genes are critical for the proper placement of segment structures of animals during early embryonic development -LRB- e.g. legs , antennae , and wings in fruit flies or the different vertebrate ribs in humans -RRB- .the diagram above shows the relationship between the genes and protein expression in flies . the protein products of hox genes are the hox proteins . hox genes are a group of related genes that determine the basic structure and orientation of an organism . interestingly , the order of the genes on the chromosome is the same as the expression of the genes in the developing embryo , with the first gene being expressed in the anterior end of the developing organism . in some organisms , especially vertebrates , the various hox genes are situated very close to one another on the chromosome in groups or clusters . in addition , homeodomains of individual hox proteins usually exhibit greater similarity to homeodomains in other species than to proteins encoded by adjacent genes within their own hox cluster . the homeodomain protein motif is highly conserved across vast evolutionary distances .

- RNN:
  hox genes are critical for the proper placement of segment structures of animals during early embryonic development lrb e g legs antennae and wings in fruit flies or the different vertebrate ribs in humans rrb. the diagram above shows the relationship the genes and protein in flies because this , it is thought hox gene clusters evolved from a single hox gene . a hox gene cluster with at least seven different hox genes was present in the common ancestor of all bilateral animals . hox genes are a group of related genes that determine the basic structure and orientation of an organism. the order of the genes is the same as the expression of the genes in the developing embryo the first gene is expressed in front of the developing organism in some organisms especially vertebrates the various hox genes are situated very close to one another on the chromosome in groups or clusters in addition homeodomains of individual hox proteins are simila in many species the homeodomain protein motif is similar in many animals

- LSTM:
  hox genes are critical for the proper placement of segment structures of animals during early embryonic development lrb e g legs antennae and wings in fruit flies or the different vertebrate ribs in humans rrb. the diagram above shows the relationship the genes and protein in flies because this , because of this it is thought hox gene clusters evolved from a single hox gene via tandem gene duplication a hox gene cluster witH seven different hox genes was probably present in the common ancestor of all bilateral animals hox genes are a group of related genes that determine the basic structure and orientation

of an organism. the order of the genes is the same as the expression of the genes in the developing embryo the first gene is expressed in front of the developing organism in some organisms especially vertebrates the various hox genes are situated very close to one another on the chromosome in groups or clusters in addition homeodomains of individual hox proteins are simila in many species the homeodomain protein motif is similar in many animals

- GRU:
  hox genes are critical for the proper placement of segment structures of animals during early embryonic development lrb e g legs antennae and wings in fruit flies or the different vertebrate ribs in humans rrb. the diagram above shows the relationship the genes and protein in flies because this , because of this it is thought hox gene clusters evolved from a single hox gene via tandem gene duplication a hox gene cluster with at least seven different hox genes was probably present in the common ancestor of all bilateral animals hox genes are a group of related genes that determine the basic structure and orientation of an organism. the order of the genes is the same as the expression of the genes in the developing embryo the first gene is expressed in front of the developing organism in some organisms especially vertebrates the various hox genes are situated very close to one another on the chromosome in groups or clusters in addition homeodomains of individual hox proteins are simila in many species the homeodomain protein motif is similar in many animals

- Transformer:
  Determine the presence or absence of certain genes or proteins that are important to the regulation of the immune response in flies, such as CDK and CDK2. Know that the hox genes are essential for the growth and function of many cells in the body and are important for many functions of the immune system. The hox genes are the most powerful genes in the Homogeneous Family of organisms (Homogeneous organisms are those that contain all or most of the hox genes). Study the expression of the genes in the developing embryo, including the anterior end of the developing organism, and the end of the developing organism. In plants and animals, the hox genes are located very close to one another on the chromosome in groups, or in clusters, i.e. Differentiate homeodomains of hox proteins from other homeodomains of hox proteins in the salivary glands of humans and animals. Recognize that the homeodomain protein motif is unique in that it is located at the very end of alladenosylmethionine (A) chains.

- Simplish.org:
  Hox genes are critical for the proper giving a place of segment structures of animals during early embryonic development -LRB- e.g. legs, antennae, and wings in fruit flies or the different vertebrate ribs in humans -RRB-.the diagram above shows the relation between the genes and protein expression in flies. the protein products of hox genes are the hox proteins. hox genes are a group of related genes that determine the basic structure and orientation of an organism. interestingly, the order of the genes on the chromosome is the same as the expression of the genes in the undergoing growth embryo, with the first gene being expressed in the anterior end of the undergoing growth organism. in some organisms, especially vertebrates, the different hox genes are placed very close to one another on the chromosome in groups or groups of things. in addition, homeodomains of individual hox proteins usually Exhibit greater likeness to homeodomains in other species than to proteins made orders for computer by adjacent genes within their own hox mass, group. the homeodomain protein pleasing design is highly kept from change across sizeable on going development distances.

- Quillbot:
  Hox genes are important for the appropriate arrangement of animal segment structures during early embryonic development -LRB-, such as legs, antennae, and wings in fruit flies, or the various vertebrate ribs in humans -RRB-. The link between genes and protein expression in flies is depicted in the diagram above. Hox proteins are the protein products of hox genes. Hox genes are a family of linked genes that govern an organism's fundamental shape and orientation. Surprisingly, the order of the genes on the chromosome corresponds to the order in which the genes are expressed in the growing embryo, with the first gene being expressed in the developing organism's anterior end. The numerous hox genes are located quite close together in several species, notably vertebrates. The numerous hox genes are located relatively close to one another on the chromosome in groups or clusters in several animals, particularly vertebrates. Furthermore, individual hox proteins' homeodomains are frequently more similar to homeodomains from other species than proteins expressed by nearby genes within their own hox cluster. Across enormous evolutionary distances, the homeodomain protein motif is remarkably conserved.