CLEF2023' SIMPLE TEXT

Darko Vujica[1], Iva Čatipović[1] and Julia Komorowska[2]

[1] *University of Split, Ruđera Boškovića 31, 21000, Split, Croatia*
[2] *University of Gdansk, Jana Bazynskiego 8, 80-309, Gdansk*

Abstract

Understanding scientific texts is crucial for success in education and beyond. However, individuals who lack expertise in a particular scientific field often encounter challenges in comprehending such documents. Scientific articles present difficulties due to their utilization of complex vocabulary, intricate language, and lengthy structures, making them inaccessible without prior knowledge. This working note presents the outcomes of SimpleText tasks 2 and 3, which involve the identification and explanation of difficult concepts, as well as the simplification of complex text passages. Various statistical and AI-based models were employed to tackle these tasks, and their performance was evaluated accordingly. Autoregressive Large Language Models (LLMs) were predominantly utilized to solve these challenges. For effective utilization of BLOOM and BLOOMZ, prompts with examples were employed, while GPT-3 demonstrated greater effectiveness with simple command prompts. In addition to LLMs, statistical and graph-based models were also incorporated into the approach.

Keywords

GPT, PKE, Wiki, SimpleT5, CLEF 2023, SimpleText, text simplification

1. Introduction

Communicating complex concepts to a general audience and simplifying scientific texts are two challenges often faced in the realm of knowledge dissemination. Explaining difficult concepts in a way that is accessible and understandable to a non-expert audience requires careful consideration of language, context, and relatability. Meanwhile, simplifying scientific texts involves condensing intricate information without compromising accuracy or losing the essence of the content. In this context, leveraging tools such as GPT-3, PKE, and SimpleT5 can aid in the process by providing assistance in generating simplified versions of scientific texts. We will write about exploring strategies for identifying and explaining difficult concepts to a general audience and delves into the simplification of scientific texts, highlighting the potential of using these advanced tools in the pursuit of effective knowledge dissemination.

[1]CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

Task 2 of the SimpleText lab competition in CLEF 2023 focused on identifying difficult concepts, explaining them, and ranking them based on their level of difficulty. To assist us in this task, we have used GPT-3 and Wiki combined with PKE. On the other hand, Task 3 aimed to generate simplified versions of scientific texts. To accomplish this, we have used GPT-3 and SimpleT5.

2. Approach

For complexity spotting, we utilized natural language processing techniques and machine learning models. We performed data preprocessing, including cleaning and tokenization. We employed term extraction methods like TF-IDF to identify important terms and assigned difficulty scores to rank them. For the identified difficult terms, we generated short explanations or definitions using rule-based techniques. Evaluation was conducted using term pooling and metrics such as accuracy and NDCG.

To simplify scientific text, we employed a combination of rule-based approaches and neural network models. Rule-based techniques were used to simplify sentence structures, replace complex terms, and simplify language. We also fine-tuned a neural network model, such as a pre-trained language model, on a large corpus of scientific and simplified text. Evaluation involved measures like SARI, ROUGE, compression ratios, and readability scores to assess the quality of simplifications.

GPT-3: GPT-3 is a powerful language model that I used to address the challenge of explaining difficult concepts and simplifying scientific texts. With GPT-3, I utilized its text generation capabilities by providing prompts and requesting simplified explanations or text versions suitable for a general audience.

PKE (Python Keyphrase Extraction): PKE played a crucial role in the process of identifying key concepts and ideas within the complex texts. By applying PKE, I extracted keyphrases from the text, which helped to pinpoint the essential terms and concepts that required simplification or explanation.

Wiki: Wikipedia served as a valuable resource for obtaining additional information and context about the identified keyphrases and concepts. I cross-referenced the keyphrases with relevant articles on Wikipedia to gather insights, examples, and concise explanations that assisted in the simplification process.

SimpleT5: SimpleT5 is a pre-trained text-to-text transfer transformer that can be fine-tuned for specific tasks such as text simplification. To simplify scientific texts, I utilized SimpleT5 by fine-tuning it on a dataset of simplified scientific texts. This allowed me to create a specialized model that generated simplified versions of scientific content with a focus on clarity and readability.

3. Results

Here is a part of the results we obtained by using GPT-3. (task 2 and 3)

Sentence: In the modern era of automation and robotics, autonomous vehicles are currently the focus of academic and industrial research.

Simplified sentence: Autonomous vehicles are being researched.

Table 1 Results

Difficult term	Difficulty	Definition
Automation	1	Automation is the use of technology to automate processes and tasks in order to increase efficiency, reduce costs, and improve quality. It involves the use of computers, robots, and other machines too.
Robotics	2	Robotics is the science and technology of designing, constructing, and operating robots. Robots are machines that are capable of carrying out a complex series of actions automatically, or with guidance.
Autonomous	2	Autonomous means being self-governing or independent. In this context, autonomous vehicles are vehicles that operate without any human input. They are capable of sensing their environment and navigating on their own.
Academic	1	Academic refers to the activities related to studying, teaching, and researching in institutions of higher education such as universities, colleges, and research institutes.
Industrial	2	Industrial: relating to or concerned with the manufacturing and production of goods by machines, factories, and other large-scale methods.

In Task 2, when comparing the performance of GPT-3 and the combination of Wiki/PKE, we observed the following results:

Sentence: With the ever increasing number of unmanned aerial vehicles getting involved in activities in the civilian and commercial domain, there is an increased need for autonomy in these systems too.

Table 2 Results of task 2 using Wiki / PKE

Difficult term	Definition

unmanned aerial vehicles	An unmanned aerial vehicle (UAV), commonly known as a drone, is an aircraft without any human pilot, crew, or passengers on board.
commercial domain	The domain name .com is a top-level domain (TLD) in the Domain Name System (DNS) of the Internet.

Table 3 Results of task 2 using GPT-3

Difficult term	Difficulty	Definition
Autonomy	2	Autonomy refers to a system's ability to operate independently and make decisions without the need for human input.
Unmanned Aerial Vehicles (UAVs)	1	Unmanned Aerial Vehicles (UAVs) are aircrafts that are operated and flown without a human pilot onboard. They are also commonly known as drones and are typically equipped with sensors.
Civilian Domain	1	Civilian domain refers to the areas of activity in which civilians, as opposed to military personnel, are involved. This includes activities such as transportation, commerce, communication, and entertainment.
Commercial Domain	2	Commercial Domain: The commercial domain, also known as the business domain, is a sector that covers the activities, strategies, and operations related to the production, distribution, and sale of goods and.
Systems Automation	2	Systems Automation is the use of technology and software to automate processes, tasks, and functions previously done manually. It can be used to increase efficiency, reduce costs, and improve accuracy.

We observed that GPT-3 excelled in identifying more difficult terms and providing detailed definitions. However, it did not perform as effectively in ranking the terms based on their level of difficulty.

4. Conclusions

The complexity spotting task aimed to identify and explain difficult concepts in scientific texts for a general audience. Our approach involved a combination of natural language processing techniques and machine learning models. By extracting important terms and assigning difficulty scores, we identified terms that required explanation and

contextualization. Short explanations or definitions were generated for these difficult terms. Evaluation using term pooling and metrics helped assess the accuracy of our approach. This task plays a crucial role in enhancing the understanding of complex scientific texts and improving accessibility for a wider audience.

The text simplification task focused on generating simplified versions of scientific texts. Our approach employed both rule-based techniques and neural network models to simplify sentence structures, replace complex terms, and make the language more accessible. By fine-tuning a neural network model on a large corpus of scientific and simplified text, we aimed to preserve the core meaning while simplifying the language. Evaluation measures such as SARI, ROUGE, compression ratios, and readability scores helped assess the quality of the simplifications. This task contributes to bridging the gap between scientific knowledge and a general audience by making complex information more easily understandable.

5. Appendix

Here are GPT-3 prompts that we have used.

GPT-3 prompt for finding difficult terms

```
def findDifficultScientificTerms(input):
    openai.api_key = "…"
    response = openai.Completion.create(
    model="text-davinci-003",
    prompt="Return a list of maximum 5 difficult scientific terms in the following
    sentence:\n\n"+input,
    temperature=0.7,
    max_tokens=40,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0
    )
    return response
```

GPT-3 prompt for rating difficult terms

```
def assign_score_for_difficult_term(sentence, term):
    import os
    import openai
    openai.api_key = …"
    response = openai.Completion.create(
    model="text-davinci-003",
    prompt="Given is the following sentence:\n"+sentence+'\nGiven is the following term   in
    the sentence: ' + term + "\nAssign a score of difficulty for the given term\
    using the following Likert scale: 1 - easy enough, 2 - difficult, 3 - very difficult. Only
    return an integer decimal number.",
    temperature=0.7,
    max_tokens=40,
    top_p=1,
    frequency_penalty=0,
```

```
    presence_penalty=0
    )
    return response
```

GPT-3 prompt for explaining difficult terms

```
def provide_definition(sentence, difficult_term):
    openai.api_key = "…"
    response = openai.Completion.create(
    model="text-davinci-003",
    prompt="Given is the following sentence: " + sentence + "\n\
    This sentence contains the following difficult term: " + difficult_term + "\n\
    Provide a definition for the difficult term.",
    temperature=0.7,
    max_tokens=40,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0
    )
    return response
```

GPT-3 prompt for simplifying sentences

```
prompt = 'Sentence: Image tampering, being readily facilitated and proliferated by today's
digital techniques, is increasingly causing problems regarding the authenticity of images.\n\
Simplification: Image tampering has become a serious problem.\n\
\n\
Sentence: Nevertheless, the interesting issue of detecting image tampering and its related
operations by using the same quantization matrix has not been fully investigated.\n\
Simplification: Methods based on image compression techniques, like quantization, are rarely
applied.\n\
\n\
Sentence: Guest virtual machines are especially vulnerable to attacks coming from their
(more privileged) host.\n\
Simplification: Guest virtual machines are vulnerable to attacks from their host.\n\
\n'

def simpleMyPrompt(prompt,input):
  import os
  import openai
  openai.api_key = "…"
  response = openai.Completion.create(
  model="text-davinci-003",
  prompt=prompt+'Sentence: '+input+'\nSimplification:', #text completion
  temperature=0.7,
  max_tokens=40,
  top_p=1,
  frequency_penalty=0,
  presence_penalty=0
)
  return response
```

6. References

[1] Liana Ermakova, Eric SanJuan, Stéphane Huet, Olivier Augereau, Hosein Azarbonyad, and Jaap Kamps. 2023. Overview of SimpleText - CLEF-2023 track on Automatic Simplification of Scientific Texts. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Anastasia Giachanou, Dan Li, Mohammad Aliannejadi, Michalis Vlachos, Guglielmo Faggioli, Nicola Ferro (Eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)

[2] Liana Ermakova, Eric SanJuan, Stéphane Huet, Olivier Augereau, Hosein Azarbonyad,
and Jaap Kamps. 2023. CLEF 2023 SimpleText Track: What Happens if General Users
Search Scientific Texts? In Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III. Springer-Verlag, Berlin, Heidelberg, 536–545. https://doi.org/10.1007/978-3-031-28241-6_62

[3] Liana Ermakova, Eric SanJuan, Jaap Kamps, Stéphane Huet, Irina Ovchinnikova, Diana
Nurbakova, Sílvia Araújo, Radia Hannachi, Elise Mathurin, and Patrice Bellot. 2022.
Overview of the CLEF 2022 SimpleText Lab: Automatic Simplification of Scientific Texts. In Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings. Springer-Verlag, Berlin, Heidelberg, 470–494. https://doi.org/10.1007/978-3-031-13643-6_28

[4] https://en.wikipedia.org/wiki/GPT-3

[5] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C Nunes, and A. Jatowt, „YAKE! Keyword extraction from single documents using multiple local features“, Informaration Sciences, vol. 509, pp. 257-289, Jan. 2020.

[6] „Hosted Inference API“ Hugging Face. https://huggingface.co/docs/api-inference/index (accessed Mai 14, 2022)

[7] https://en.wikipedia.org/wiki/Text_simplification

[8] O. M. Cumbicus-Pineda, I. Gonzalez-Dios, and A. Soroa, ‘A Syntax-Aware Edit-based System for Text Simplification’, International Conference Recent Advances in Natural Language Processing, RANLP, 2021, pp. 324–334.