CLEF SimpleText Lab

Nick Largey, Finn Michaud, Ben Gaudreau, and Gabrielle Akers

What is SimpleText Task 1?

CLEF 2023 SimpleText's task 1 asks us the question "What is in (or out)?"

The goal of SimpleText Task 1 is to evaluate given articles to determine how based on relevant they are in relation to a query, and extract any relevant passages. These passages will be rated based on their text complexity and how credible they are. This information will be used in future tasks to determine which articles need to be simplified.

All articles will be rated between 0 and 2 for each area, with 0 being easy to understand or not relevant and 2 being complex or relevant.

Related Work

- University of Amsterdam have done this experiment using both a BERT cross-encoder and comparing with the elasticsearch as their baseline. One BERT model run did the top 100 and the other doing the top 1000 in relevance. Elasticsearch was used to calculate the credibility and readability.
- Another group taking part in the CLEF 2023 SimpleText lab also tested using a bi-encoder as well as a cross-encoder. They found that using a GPL was the most efficient method.
- Some related work outside of CLEF went into studying retrieving information using a DPR to improve the performance by working alongside a couple of independent BERT networks.

Our Approach

Our approach to this task is to use a cross-encoder to select relevant passages from the articles.

After finding all relevant passages we intend on using a bi-encoder BERT model, alongside a DPR to classify the data in terms of how credible and how complex they are. This will allow us to help counteract some of the limitations of the BERT model.

Task 2 - The Goal

- “Identifying and explaining difficult concepts”
- Which concepts in scientific abstracts require more context and explanation in order to help a reader understand a text better
- Complexity spotting - Up to 5 terms in a given passage should be identified, ranked, and a meaningful definition/explanation should be provided – no paraphrasing!
- Complexity can be based off different types of indexes such as the Flesch-Kincaid grade level (what grade a person who have to be in the US education system to understand the text). However, SimpleText doesn’t supply a particular index.

Related Work (CLEF 2023 Task 2)

- University of Guayaquil/Jaen used zero-shot and few-shot learning techniques on GPT-3 with “prompt engineering”
- University of Southern Maine used keyword extraction approaches with YAKE and KABIR along with IDF weighting for complexity/ranking. For defining they used top-ranked documents derived from a trained classifier.
- University of Split used GPT-3 and TF-IDF for complexity spotting and for definitions they used Wikipedia with GPT-3 to handle explanations.

The Approach (may change)

- Develop a pipeline -> complexity spotting, ranking, and defining
- Interested in using YAKE or keyBERT for complexity spotting because of their effectiveness at key term extraction (part of the complexity spotting process)
- Interested in using simpleT5 or BLOOMZ for generating definitions/explanations for complex terms because of past CLEF members having success with these.

What is SimpleText Task 3?

Task 3 - Simplifying Scientific Text - The essence of the entire SimpleText lab

- Learning from and working in combination with my other 3 team members, my mission for Task 3 is to read in an entire scientific text as input, and create a simplification of that text, so that a commoner, with no extensive scientific training, will be able to understand the broad scope of what the text covers.

Related Works

- Many of the past submissions from 2022 and 2023 utilized similar approaches where the teams would use a version of GPT to prompt engineer an initial simplification of the input text, then use a T5 model to generate the final output.
- One team from E.T.S.I. Informática in Spain used the COVID-SciBERT model as a masking layer to identify sentences that they considered to be “simple concepts” so that they wouldn’t need to simplify them in their more costly T5 model.
- Most team’s scores clustered around the 8 - 14 mark for the Flesch-Kincaid grade level (FKGL) - a metric for measuring the readers comprehension level (the lower the score, the younger the reader) and around the 35 - 45 mark for the SARI (System output Against References and against the Input sentence), which is a metric for measuring the quality of words added, deleted or kept to gauge the quality of the simplification, mainly in terms of preserved meaning and improved readability.

Planned Approach

Working with my teammates, I plan to learn from their successes and struggles in order to understand how to best approach the use of Llama-2, Orca and ScispaCy. Both Llama-2 and Orca are open source LLM's so my hope is that I will be able to gain a better understanding of how to manage changes rather than blindly experimenting with input changes in a closed source model like ChatGPT.