Language Agnostic News Summarization (LANS)

Gaurav Aggarwal
Stony Brook University
gaggarwal@cs.stonybrook.edu

Jayarama Das Krovvidi
Stony Brook University
jkrovvidi@cs.stonybrook.edu

Shreejay V Jahagirdar
Stony Brook University
sjahagirdar@cs.stonybrook.edu

Abstract

The amount of data available on the internet is vast, and sometimes it is practically impossible to review the entire information thoroughly. Understanding the abstract idea of what the information wants to convey is beneficial before going through the details. We construct a model that takes a news article in any language(currently supporting French) with an existing headline and generates our summary version, making the system language independent, and will be helpful in the analysis and comparison of the similarities, relevance, and evaluate any semantic inconsistency/bias with the original headline.

1 Introduction

Summaries are helpful in the analysis and comparison of the similarities and relevance with the original headline and can help the user understand whether the report or story is relevant to what they are looking for.

There are two broad strategies for summarization, extractive, and abstractive summarization. While the first deals with scoring, ranking, and extracting the most prominent sentences from the input text, the latter functions more like a human and tries to understand the context before generating the summary.

The challenge with text summarization is the complication of human language and how humans express themselves. Text can vary in length, using adjectives, idioms, and general knowledge to make the text expressive, making the text summarization even harder for the machines. Human text generally contains pronouns to address the entities, and understanding which pronoun replaces which entity referred to before is called an "anaphora problem" and complicates the task.

Summary evaluation is an even more complex problem. If you are to believe that the summary is a reliable standby for the source, you must be self-assured that it reflects precisely what is related to the primary source. Therefore, approaches for generating and evaluating summaries must accompany each other.

The issue of semantic inconsistencies can be handled by a summarization engine that is fine-tuned on relevant examples and one that can effectively capture context over multiple paragraphs of text. One of the key ideas while addressing this issue was to use a context-aware transformer architecture rather than a bi-directional RNN, which should improve performance in the long run. The main motivation while evaluating the model on downstream semantic analysis is to use metrics like ROGUE that can effectively compare the semantic similarity instead of direct word-to-word matching.

Some of the key outcomes that were pivotal for the project are as follows

1. We implemented a transformer-based summarization engine for summarizing news articles.

2. We developed a translation module to address the problem of summarizing text from different languages.

3. Our evaluation clearly shows an existing distinction between the semantics of an article with its existing headline.

2 Task

The workflow (Figure 1) - api takes a text and its heading in French, which is passed through a translator to translate the text into English. The translated text is fed into a trained summarizer model, which summarizes it into an English heading. This predicted heading is then translated back into the source language, and finally, a similarity score is generated, which depicts the meaningfulness of the article's headline.

Current State-of-the-art solutions use a similar method with a transformer-based architecture trained in a plethora of languages like XLM and RoBerta [1]

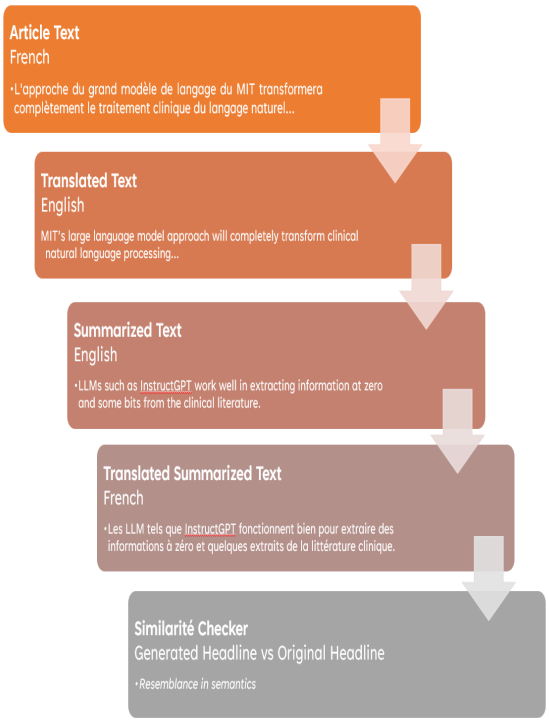

Figure 1: Workflow

2.1 Baseline Model(s)

We used a transformer-based approach and used RoBERTa [2] and T5 [3] models as our baseline. We couldn't use an LSTM-based model due to slow training time.

RoBERTa is an optimized BERT model, which was heavily under-trained with an inefficient training approach. RoBERTa provided efficient, longer training on the BERT architecture to produce state-of-the-art results in almost all Language Modelling tasks.

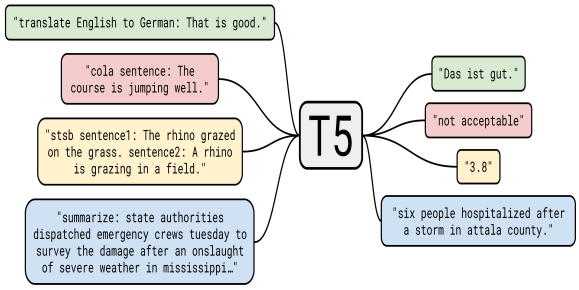

Figure 2: T5 Model

BERT architecture (Figure 3) uses an attention mechanism that learns the contextual relations between words in a text. Transformers include two separate tools — an encoder that reads the text input and encodes it into an encoded representation and a decoder that produces a prediction for the task based on constructed encoded representation [4].

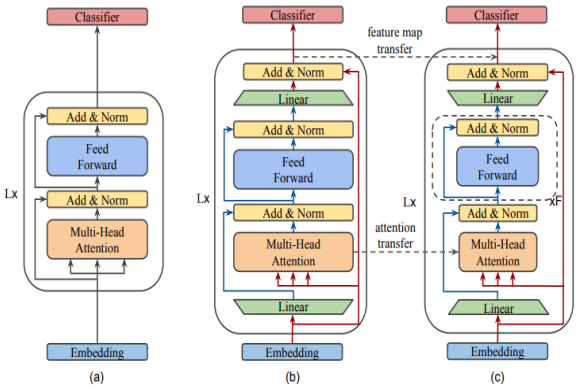

Figure 3: BERT Architecture

This mechanism can be used separately for various tasks, mainly of the encoder (Figure 4) - to produce an encoded representation of the sentence, find similarities between two texts, etc., which we explored in this project. The encoded representation is the last hidden layer output of the encoder architecture, which is further used by the decoder system to produce the required Language Modelling task.

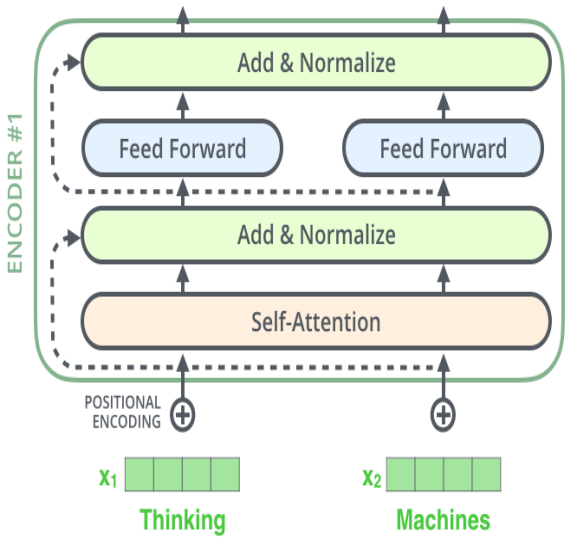

Figure 4: Transformer Encoder

2.2 The Issues

Since RoBERTa is trained on Masked Language Modelling (MLM), it performs well on Natural Language Understanding (NLU) tasks like classification, regression, etc., but struggles in language generation tasks like summarization, translation, etc., which are required in our project.

Since T5 works better on a variety of tasks, we made it our baseline for further finetuning the model on our dataset.

3 Approach

Since we wanted to make a language-agnostic system, we used the concept of pivoting for text summarization. We started with the text, which can be any language (currently supporting only french), translated into English, and then proceeded with the summary. We selected English as the pivot language because of the sheer amount of English news dataset present, which made the training summarizer model easy.

For the translation, we used facebook's BART pre-trained model, which is trained to support 50 language pairs of translation. We used the model to translate the text from french to English and translate the predicted heading back to french for a similarity score. T5 also supports language translation, but due to the following reasons, we decided to move ahead with BART.

- Since we are using t5 for the summarization, we wanted to use a different model for translation. According to our hypothesis, using the same model could include an inherent bias in the encodings.
- BART supports more than 50 language pairs for the translation task, which made the system language independent for future language support.

The next step required the translated English text and fed into our fine-tuned t5-small model to generate a text heading (summary) in English. The English summary is then translated back to the source language using the same translator.

Source translated heading along with the original heading is then provided as input to a function that calculates the similarity score between the two. A higher similarity score means the title captures the context and is relevant to the text. A lower score means it doesn't capture the correct context or is most likely a clickbait title.

3.1 Similarity and Resemblance

To give a concise and correct heading for any article, we trained our model to summarize a given English text. We then used a method to determine the similarity between the original title and the generated output. If they are similar, then we can assume that the original heading is well-written and otherwise not contains a click-bait title.

3.2 Exposing Bias

In general, many news articles have an inherent bias toward a community, race, gender, or class of objects according to the demographic of the source. The project explores this bias through a WEAT test in the source language.

3.3 Implementation Details

After reviewing the related work on text summarization, we decided to write scripts for importing the data into a preferable 'Article' format after scraping through millions of review articles online from Google's NewSHead dataset.

```
{
   "urls":
   [
      "https://golfweek.com/2019/05/13/
          tiger-woods-faces-wrongful-
          death-lawsuit-brought-by-
          parents-of-restaurant-employee
          /",
      "https://www.nbcnews.com/news/us-
          news/tiger-woods-sued-drunk-
          driving-death-former-employee-
          n1005251",
      "https://www.wpbf.com/article/
          local-parents-sue-tiger-woods-
          for-wrongful-death-of-their-
          son/27460400",
      "https://www.cnn.com/2019/05/13/us
          /tiger-woods-wrongful-death-
          lawsuit/index.html",
      "https://www.usatoday.com/story/
          sports/golf/2019/05/13/tiger-
          woods-wrongful-death-lawsuit-
          florida/1195131001/"
   ],
   "label": "Tiger Woods wrongful death
       lawsuit."
}
```


We trained our baseline models with this data, but since it contained a lot of noise, the results were unsatisfactory. We then switched to a translate-summarize-similarity architecture, as explained above.

For the translation, we used a pre-trained BART model, with default tokenizer and model weights, for the French language.

We started with the t5-large and t5-base models for the primary summarizer model. But due to resource constraints, exceptionally high memory requirements, and giant model sizes, we could only train the t5-small model due to a lack of resources. We fine-tuned the t5-small with approximately 70 million parameters on the CNN-daily news dataset for the specific summary generation task.

We used a sample of 10,000 records to prototype the approach for two epochs and 8 batch sizes, which took approximately 1.5 hours on GPU. After satisfactory results, we took a large sample of 50,000 training data points and 5,000 for each validation and test data point. We trained our model for ten epochs with a batch size of 16, which took roughly 11 hours.


```
"hyperparameters":
{
   "epochs": 10,
   "batch_size": 16,
   "learning_rate": 2e-5,
   "weight_decay": 0.01,
   "encoding_length": 128
}
```


We evaluated this model on the ROUGE scores on the test data to evaluate the model, finally saving the model for further tasks.

This saved model made the base for the summary task and provided a summary of any given news article. We used this heading and translated it back to the source language.

We used the two headings to calculate the similarity scores and extracted the embedding representation using the RoBERTa model from its last hidden state output. These are used to calculate a semantic cosine similarity instead of a word-to-word similarity of the sentences, which provides a more robust approach.

We also evaluated the bias of these embeddings on some pre-defined classes like gender, sports, and other news-dominated subjects.

4 Evaluation

To test the similarity between the generated summary and pre-existing headline or title, we need some metric that can evaluate based on their sentence representations. This metric will also be used in training our summarization model for calculating loss for gradients. We will use the following metrics, which help evaluate sentences: Loss, ROUGE1 - a similarity metric for unigrams, and ROUGE-L - a similarity metric for Longest Common Subsequences.

4.1 Dataset Details

We tried using *Google's NewSHead*[5] dataset, which contains over 350,000 records of news URLs from different sources and domains, and made several scripts to extract the news articles. We trained the model on the data set, but the article parsing added too much noise making the data unusable.

We then used the *cnn-dailymail* [6] dataset to train our summarizer model, which is a collection of

287,000 indexed records in English with text, title, and author details. We made a (50,000 – 5,000 – 5,000) split on the articles' data into train, validation, and test sets.

Text	Heading
LONDON, England (Reuters) – Harry Potter star Daniel Radcliffe gains access to a reported £20 million ($41.1 million) fortune as he turns 18 on Monday, but he insists the money won't cast a spell on him. Daniel Radcliffe as Harry Potter in "Harry Potter and the Order of the Phoenix"...	"Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday. The young actor says he has no plans to fritter his cash away. Radcliffe's earnings from the first five Potter films have been held in the trust fund ."

Table 1: CNN Daily News Dataset

For the end-to-end translation-summarization-similarity testing, we used the *mlSum* [7] dataset, which contains 1.5M+ articles in 5 languages. We used only the french data points and included the main text, title, and author details.

4.2 Evaluation Measures

The evaluation measure includes the loss, ROUGE-1, and ROUGE-L scores which are presented in the following table (Table 2).

Epoch	Loss	ROUGE-1	ROUGE-L
2	2.428	23.4339	19.2173
4	2.416	23.4291	19.2041
6	2.410	23.4008	19.1982
8	2.407	23.3984	19.1945
10	2.405	23.3876	19.1937

Table 2: Evaluation Metrics

The loss v/s epochs are plotted to see and evaluate how the loss value is decreased sequentially (Figure 5).

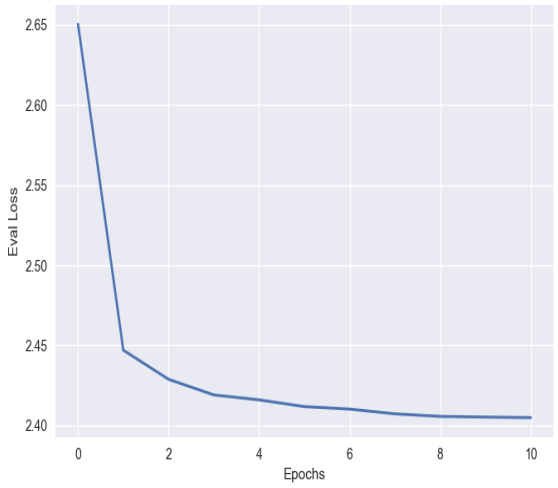

Figure 5: Loss

Similarly, Rouge1 v/s epochs are plotted to see and evaluate how the ROUGE score are varying with the training (Figure 6).

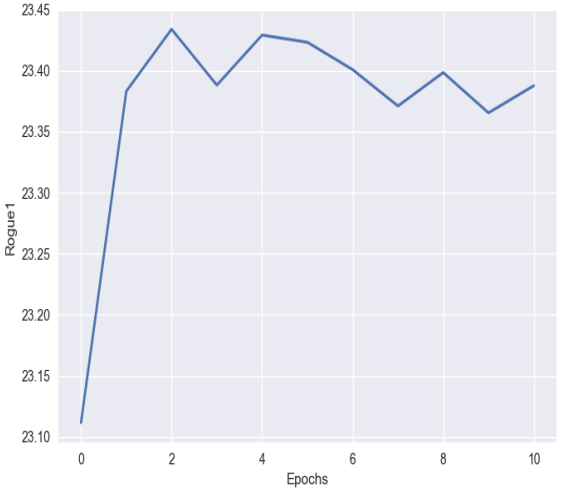

Figure 6: ROUGE-1

4.3 Results

We calculated the similarity scores on the 500 training random samples from the *mlm* dataset and plotted the similarity of the articles with their original headings (Figure 7).

We found that most of the articles were having similarities between the 40-60% range. This proves our hypothesis that almost half of the articles out there have click-bait tiles as per our

results in the era of the global nature of these articles.

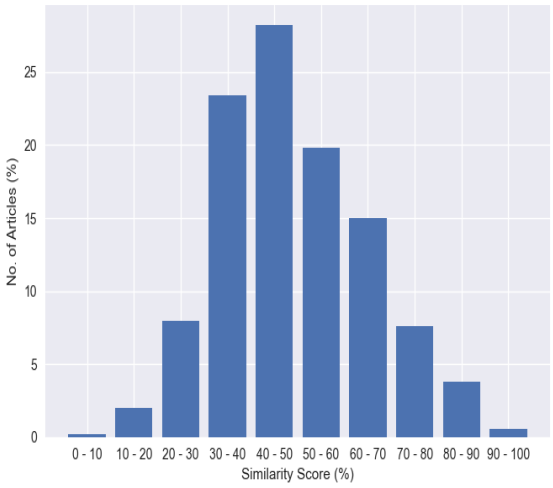

Figure 7: Scores

4.4 Analysis

To expose the bias in the articles, we performed and analyzed the Word Embedding Association Test (WEAT) test on the dataset. Since the data comes from the French demographic, the scores aligned with the expectations.

The results are as follows:

```
{
    White_Black_Pleasant_Unpleasant:
        -0.22020559,
    Male_Female_Career_Family:
        0.75734985,
    Math_Arts_Male_Female: 0.65289456,
    French_US_Pleasant_Unpleasant:
        -0.06346025,
    Football_Golf_Popular_Unpopular:
        0.6002448
}
```

Male - Female - gender bias showed a strong score which is inherently present in every news channel. The same can be expected with the race and sports classes which aligns with the general view of the people. Ahead of UEFA Euro 2016 front pages in French newspaper dominated of racism due to lack of players of North African origins in the team. While the true reasons remain a mystery people sometimes do say that a racial bias exists. Our analysis also shows somewhat similar results depicting a bias is present, which is what we intended to find out as one of our use case of our system.

1. The thing that worked is our hypothesis that most of the articles have click-bait titles. It is confirmed by the similarity scores of the data points.

2. The architecture to translate the language into pivot language worked properly while providing a system-independent structure.

3. As we can observe, the model didn't decrease the loss much after epochs 6-7, and poor ROUGE scores compared to the state of art models.

4. This can be due to the already heavily trained T5 model, and further learning didn't bring much of a difference. Also, the entire pipeline can be efficiently used to avoid the need for multiple separate systems.

4.5 Code

Link to the Codebase along with README and model is provided here.

5 Conclusions

There are always some human-induced inaccuracies in news article headlines. Still, we provide a way for readers to skim through these titles effectively to save time and make the overall experience more engaging.

During the project, we understood the importance of choosing the suitable model and parameters experimentation based on the downstream task, like the difference between selecting an encoder-decoder transformer model that works on LSTMs vs. Transformers.

Data noise is important in any NLP project and should be dealt with honestly.

6 Future Work

- We can extend this project to support more languages.

- Due to resource constraints, we could not explore the larger model architecture, which can be explored further.

- We trained the model on a small sample (1/5th) of the data. We can train for longer periods of time.

- The project can be extended with the functionality of a web crawler, classifying the news articles as spam/non-spam based on headline correctness.

7 Related Work

[8] This paper presents an abstractive summarization technique using Attentional EncoderDecoder Recurrent Neural Networks and proposes novel models to address critical problems such as emitting rare or unseen words at training time.

[9] Discusses the importance of summarization and proposes an improved strategy by combining the TextRank and BART models to solve the deviation problem. The outputs from both these models are weaved together to augment the weights of pivotal sentences in the input, making it more thematic. This combined output is passed through the BERT model to get the final summarization.

[10] Explains the challenges of automatic summarization and the importance of strategies that generate a compressed interpretation of the input. These strategies might use words that may not be found in the original input.

References

[1] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[5] Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, Hongkun Yu, You Wu, Cong Yu, Daniel Finnie, Jiaqi Zhai, and Nicholas Zukoski. Generating Representative Headlines for News Stories. In *Proc. of the the Web Conf. 2020*, 2020.

[6] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701, 2015.

[7] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. Mlsum: The multilingual summarization corpus. *arXiv preprint arXiv:2004.14900*, 2020.

[8] Ramesh Nallapati, Bowen Zhou, Dos Cicero, Santos, Çaglar Gulçehre, and Bing Xiang. *Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond*. Aug 2016.

[9] Yisong Chen and Qing Song. News text summarization method based on bart-textrank model. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 5, pages 2005–2010, 2021.

[10] Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66, Mar 2016.