Final Graded Assignment

NLP

UNIVERSITY OF TÜBINGEN, 2024 SUMMER

Ece Sena Etoglu

August 13, 2024

1 High-Level Index of the Report

1.1 Dataset Analysis

Explore and understand the dataset, including its key characteristics and structure.

1.2 Part 1: LDA Model Development

- **Preprocess the Dataset**: Prepare the data for LDA model training.
- **Conduct Experiments**: Perform preliminary experiments and visualize the LDA model to uncover the main topics.

1.3 Part 2: Model Evaluation

- **Evaluate T5 and BART Models**: Conduct experiments with T5 and BART on the summarization task using the EdinburghNLP/xsum dataset.
- **Performance Metrics**: Assess model performance using ROUGE, BERTScore, METEOR, and Human Evaluation.
- **Conclusions**: Summarize the findings and provide a final conclusion.

2 Dataset Analysis

The dataset used is the `EdinburghNLP/xsum` dataset, which is a well-regarded benchmark for abstractive summarization tasks. It can be accessed at `https://huggingface.co/datasets/EdinburghNLP/xsum`.

2.1 Data Loading and Preparation

Due to limited computational resources, the dataset has been truncated. To ensure randomness, the dataset was shuffled before truncation.

2.2 Dataset Characteristics

- **Training Set**: 15,000 samples
- **Validation Set**: 3,000 samples
- **Test Set**: 3,000 samples
- **Summary Length**: Each summary is a single sentence, indicative of abstract summarization.

2.3 Observations

- The dataset is clean, with each summary conforming to the one-sentence format.
- Despite the brevity of summaries, they exhibit similar word and sentence densities compared to the full documents.
- Full documents contain more punctuation and stopwords, reflecting higher complexity.
- Word clouds indicate that summaries deliver a concise and focused description of the articles.

3 Creating an LDA Model

3.1 Introduction

Latent Dirichlet Allocation (LDA) is a statistical model used for topic modeling and text mining. The core idea is to infer the latent (hidden) topic structure that best explains the observed words in the documents. Preprocessing is crucial before feeding data into the LDA to come up with better representations.

3.2 Implementation

3.2.1 Preprocessing Pipeline for LDA

Standardization:

- Convert the text to lowercase.
- Remove newline characters and excess whitespace.

Tokenization:

- Split the text into individual words (tokens).
- Simple “word_tokenize” is used. Although there are more advanced tokenizers, “word_tokenize” is the most common approach in LDA representations.

Removal of Non-Alphabetic Tokens:

- Filter out tokens that are not alphabetic.

Stopword Removal:

- Remove common stopwords (e.g., “the”, “and”) to focus on meaningful words.

Part-of-Speech (POS) Tagging and Lemmatization:

- Apply POS tagging to assign grammatical tags to each token.
- Lemmatize tokens based on their POS tags to reduce them to their base forms.

Bigram and Trigram Detection:

- Identify bigrams (two-word phrases) and trigrams (three-word phrases) in the text to capture multi-word expressions.

Dictionary Creation:

- Create a dictionary from the preprocessed texts, filtering out words that appear too infrequently or too frequently.

Bag-of-Words (BoW) Model:

- Convert the preprocessed texts into a Bag-of-Words (BoW) representation, where each text is represented as a vector of word counts.

3.2.2 Experiments to find the optimum topic number

Coherence is a metric used to evaluate the quality of topics generated by LDA models. It measures the degree of semantic similarity between the top words in each topic. A higher coherence score indicates that the words within a topic are more related to each other, suggesting that the topics are more meaningful and interpretable. By experimenting with different numbers of topics and observing their coherence scores, one can identify the optimal number of topics that yields the most coherent and useful topics for the given dataset.

Plot below represents coherence - topic number plot of LDA models that only differ in topic number.

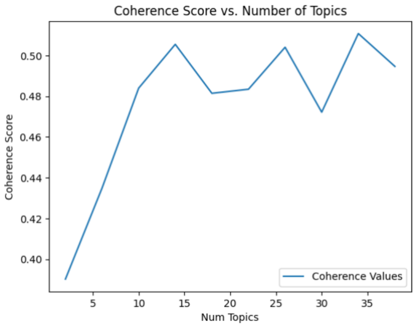


Figure 1: Coherence vs. Topic Number plot for LDA models with varying topic numbers. Topic 14 has the best coherence score. Although there is another peak at topic number 34, it is in a portion of the graph that is not strictly ascending, and its visualization is not optimal.

For other experiments, bad performing LDA topics and details such as visualizations please refer to the .ipynb notebook. In the followibg results section, best performing model's results are displayed

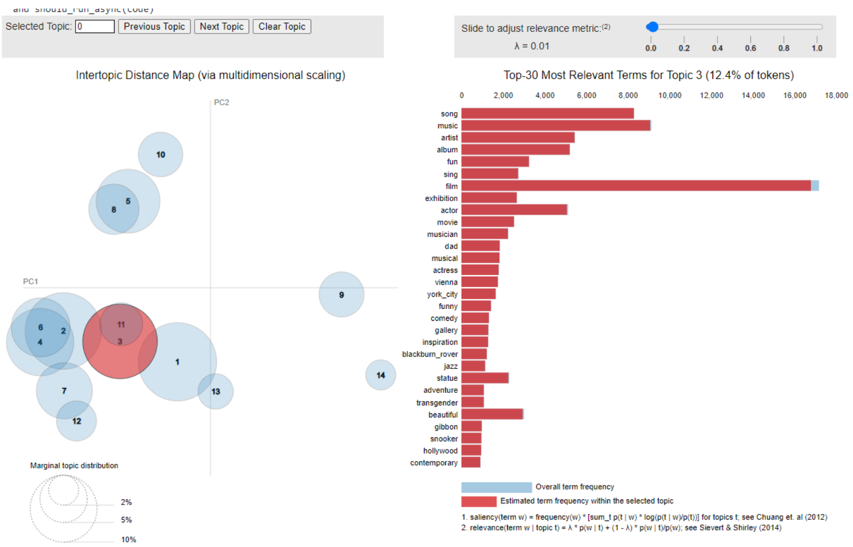
3.3 How to Visualize LDA Models

pyLDAvis library provides highly useful visualizations.

3.4 How to Evaluate LDA Models using pyLDAvis Library

- Avoid Overlapping Topics: Topics should be well-separated with minimal overlap. Overlapping circles suggest that the topics are too similar or indistinct.
- Avoid Scattered Topics: Topics should not be excessively scattered. Scattered topics indicate that the model might be too granular or not capturing coherent themes.

3.5 Results and Discussion



```
and should_run_async(code)
```

Selected Topic: 0

Slide to adjust relevance metric:(2)
λ = 0.01

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 2: It can be seen that the data has less overlesss and is distributed better in the space. But these observations are not enough, also words that are grouped together should lead to a coherent structure.
There are 14 topics detected by the model. The below topics are coherent accord. to the keywords indicated by the model:

- 1 - GENERAL SPORTS
- 2- POLITICS
- 3- ENTERTAINMENT
- 5- CRIME
- 6- INTERNATIONAL POLITICS
- 7- BUSINESS & ECONOMY
- 8 - INTERNATIONAL CRIME / TERROR
- 9 - FOOTBALL
- 12- SCIENCE
- 13- INTERNATIONAL SPORTS (olimpycs)
- 14- WORLD NEWS

The word distribution of the topics that I didn't label was not as coherent as the other ones, a more complex model i.e with higher iterations can be trained to get better topic representations. Due to the limited sources, this model was trained with 400 iterations and 10 passes

4 Comparison of T5 and BART on Abstractive Text Summarization

4.1 Introduction

- **What is Abstractive Summarization?:**
  Abstractive summarization is a task where the system generates a concise summary by understanding the content and rephrasing it, rather than only extracting sentences from the original text. This approach involves generating new phrases and sentences that convey the main ideas of the source material, similar to how a human would summarize text.

- **Architectures of the models T5 and BART:**

  T5 (Text-to-Text Transfer Transformer) utilizes a vanilla encoder-decoder transformer architecture. It is pre-trained using a span corruption objective known as *text infilling*. In this approach, contiguous spans of text are masked out, and the model learns to predict these masked spans from the surrounding context. This method helps T5 learn to generate missing segments of text based on the context provided by the unmasked portions.

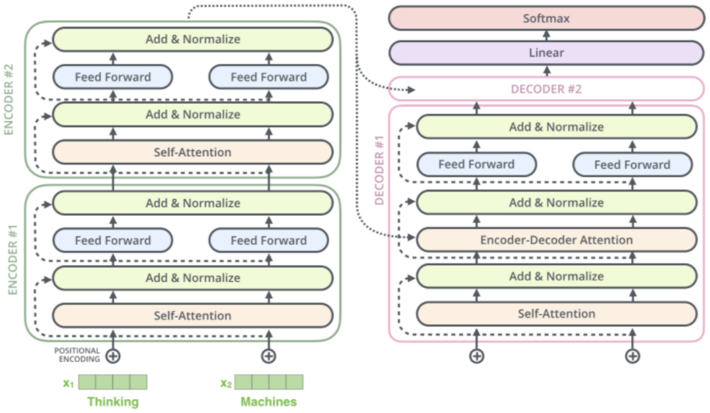


Figure 3: Architecture of T5 (Text-to-Text Transfer Transformer).

BART (Bidirectional and Auto-Regressive Transformers) is a sequence-to-sequence (Seq2Seq) model that combines elements of BERT and GPT. Its architecture includes a bidirectional encoder, similar to BERT, and an autoregressive decoder, similar to GPT. BART is pretrained using a denoising autoencoder objective. This involves corrupting the input text in various ways, such as token masking, sentence permutation, and document rotation. The model is then trained to reconstruct the original from these corrupted inputs.

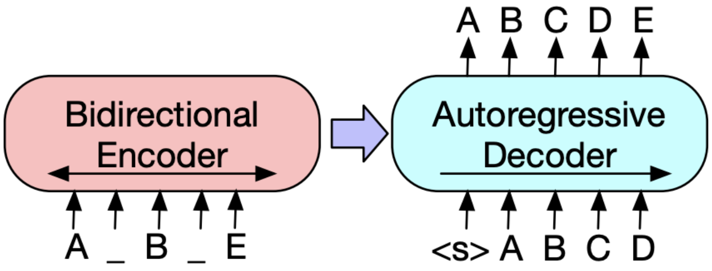


Figure 4: Architecture of BART (Bidirectional and Auto-Regressive Transformers).

Similarities and Differences

Both T5 and BART employ corruption and reconstruction techniques to pretrain their models, However, there are differences:

  - **Training Approach:** BART's pretraining involves a broader range of corruption techniques, including sentence shuffling and document rotation, alongside token masking. This diversity helps the model handle a wide variety of distortions. T5 specifically uses span corruption (text infilling), focusing on predicting masked-out spans within the text. This approach is only targeted at learning to fill in missing segments of text.
  - **Architecture:** T5 uses a traditional vanilla encoder-decoder transformer, whereas BART uses a bidirectional encoder and an autoregressive decoder.

- **Metrics to Evaluate: ROUGE, BERTScore, METEOR, and Human Evaluation:**

  ROUGE is a set of metrics that compare the overlap of n-grams between the generated summary and reference summaries.

  BERTScore uses BERT embeddings to compare the similarity of tokens between the generated and reference summaries, providing a more semantic evaluation.

  METEOR evaluates based on the alignment of chunks between the summary and references, considering synonyms and stemming also provides a more semantic evaluation.

  Human Evaluation it is crucial for humans to evaluate outputs. I evaluated some of the sentences myself.

- **Importance of Decoding Techniques: Temperature, Top-p, and Top-k:**
  Decoding techniques like temperature, top-p (nucleus sampling), and top-k sampling are crucial for controlling the diversity, quality of generated text and to reduce repetition.:
  Temperature controls randomness, where where higher values introduce more diverse results.
  Top-k limits the sampling pool to the top k most likely tokens, while **top-p** includes the smallest possible set of tokens whose cumulative probability exceeds a threshold, ensuring a balance between diversity and coherence.

4.2 Experiments Pipeline

Below is the high-level pipeline used for conducting the experiments. Details for each section are included in the .ipynb notebook.

4.2.1 Preprocessing of the Dataset

The transformer models T5 and BART come with their own tokenizers, which handle the tokenization of the input text. The dataset was already clean, therefore, additional preprocessing steps such as tokenization are not necessary before feeding the data into these models.

And in fact, in most of the examples it was not done as it might lead to performance reduction.

These models convert the tokenized input directly into embeddings, which are then used for further processing in the model's architecture.

Hence only the tokenizers of transformers are used to preprocess the data. Minor adjustments that were made:

- T5:
  It is standard practice to include a prefix when using T5 for specific tasks to guide the model appropriately. For the summarization task, the prefix "summarize:" was added to the input data, as commonly recommended in T5 implementations.
- BART:
  There are mixed practices regarding the use of a prefix with BART. To explore its impact, two variations of the dataset were prepared: one with a summarization prefix and one without. The results of both approaches were then evaluated on the validation set.

4.2.2 Parameter Tuning for Models to Find the Best Parameters

1. Generate summaries on the validation set using different combinations of top-p, top-k, and temperature.
2. Record the results for each experiment and print a subset of summaries for that experiment.
3. Display the results using ROUGE scores (a common metric for summarization, chosen for its simplicity).
4. Pick the best performing experiment by prioritizing the ROUGE-2 score.

4.2.3 Generating Summaries on the Test Set Using Parameters Found Above

1. Generate summaries using the best experiment obtained from parameter tuning.

4.2.4 Comparing Two Models with Various Metrics

1. Compare models according to the following metrics:
    (a) ROUGE Scores
    (b) BERT Scores
    (c) Human Evaluation (my evaluation)
    (d) METEOR Score

4.3 Results and Discussion

- **Preprocessing BART, adding a prefix or not?**

  For BART, an issue was observed when including the prefix "summarize:". The model unexpectedly included the prefix as the starting token in the generated summaries. Also overall reduction in output quality was observed. Therefore it can be concluded that the prefix "summarize:" should not be added to BART. Outputs with and without this prefix can be observed in the notebook.

- **ROUGE Scores**

Metric	T5	BART
ROUGE-1	0.2025	0.1583
ROUGE-2	0.0316	0.0289
ROUGE-L	0.1375	0.1008

Table 1: ROUGE scores for T5 and BART models.

  ROUGE scores are indicators of content overlap. Since T5 consistently outperforms BART in all ROUGE metrics, it indicates that T5 generates summaries with more content coverage and alignment with the reference summary. If performance varied, ROUGE-2 would be the metric of interest due to its focus on capturing key phrases and relationships.

  Since abstractive summarization is an open ended task. Content overlap is not enough and more metrics are needed:

- **BERTScore**

Model	Average Precision	Average Recall	Average F1 Score
BART	0.8201	0.8769	0.8475
T5	0.8453	0.8654	0.8551

Table 2: BERTScore metrics for T5 and BART models.

For abstractive summarization, a balance between relevance (precision) and comprehensiveness (recall) is desired. The F1 score provides a balanced view of how well the summary maintains both high relevance and broad coverage of key information. T5 performs better according to the F1 score of BERTScore, indicating that T5's summaries are more semantically accurate and align better with the reference summaries.

- **METEOR Score**

Model	Average METEOR Score
BART	0.2176
T5	0.1867

Table 3: METEOR scores for T5 and BART models.

BART outperforms T5 in METEOR, suggesting that BART generates summaries that are more effectively aligned with reference summaries in terms of word choices, order, and semantic variations.The comperatibly lower meteor score of T5 could be due to the difference in their pretraining objectives:
BART uses a more complexed span corruption that includes various span corruption techniques. Making it better at of word choices, order, and semantic variations comparing to T5.

- **Human Evaluation:**

I wanted to evaluate generated summaries just by reading.To evaluate the generated summaries effectively, I used reference summaries as a benchmark. Direct comparison with the original documents would be impractical due to their length and complexity. The reference summaries, although concise, capture the essence of the key information. Below are examples from generated summaries by BART and T5:

Example 1 from Notebook:

Reference Summary:
"A pay rise "significantly higher than 1%" will be required in future to ensure an "adequate supply of good teachers" in England and Wales, the School Teachers' Review Body warns."

BART Summary:
"It says increases will be needed if current trends in teacher recruitment and retention continue.Nevertheless, the STRB has recommended a pay

rise of just 1% for teachers in England and Wales from this September.The rise applies to all leadership and teacher pay ranges and to allowances.Image copyright Getty ImagesTeachers' unions expressed disappointment at the 1% rise, saying it would do nothing to ease recruitment or retention issues.The STRB report said: "For September 2016 we [...] recommend a 1% uplift to the minima and maxima of all classroom teacher pay range and leadership pay ranges, to all teacher allowances and teacher allowances, to classroom teacher allowances in the national pay framework, and to teacher allowances."The report said the rise would"

T5 Summary:
"the STRB recommends a pay rise of just 1% for teachers in england and wales. the rise applies to all leadership and teacher pay ranges and to allowances. teachers' unions expressed disappointment at the recommended 1% rise."

Comparison:

- **Completeness:** BART provides a thorough explanation, including details about the context and implications of the recommended pay rise. However, it includes extraneous details and is somewhat lengthy, with some parts being cut off even though the models were initialized with same number of max sequence tokens. T5, on the other hand, is more concise and focuses on the essential information, though it omits some of the contextual details present in the reference summary.
- **Clarity:** BART's summary is detailed but lengthy, which may impact readability and clarity. T5's summary is succinct and straightforward, making it easier to grasp the main points quickly but without the added context that BART includes.

Example 3 from Notebook:

Reference Summary:
"The Welsh Rugby Union (WRU) takeover of Newport Gwent Dragons will be put to the vote on Tuesday, 9 May."

BART Summary:
"Media playback is not supported on this device Newport RFC's board approves takeover of Rodney ParadeThe takeover, which would include the Rodney Parade ground, has already been approved by the WRU and Dragons boards.It needs to be ratified by 75 per cent of Newport RFC shareholders next month.Dragons chief executive Stuart Davies agreed the decision was pivotal, adding: "It is a hugely important moment in the history of Newport rugby, the Dragons and Rodney Parade."He warned that holding company Rodney Parade Ltd faced going into liquidation of a deal was not agreed.The shareholders have been issued with a document

providing information on the central issues.It also includes an open letter from Tony Brown, one of the two main financial backers, advocating the deal."

T5 Summary:
"the takeover, which would include the Rodney Parade ground, has already been approved by the WRU and Dragons boards. it needs to be ratified by 75 per cent of Newport RFC shareholders next month. the ground is used by the Dragons regional side, Newport RFC and League Two football club Newport County."

Comparison:

- **Completeness:** BART offers a comprehensive summary with detailed background information, including the historical significance of the event and reactions from key figures. But again, is cut off. T5 is more concise, covering the essential points of the takeover but lacking the additional context and reactions provided by BART.
- **Clarity:** BART's summary, while detailed, can be overwhelming due to its length and the amount of information provided. T5 is clear and to the point, making it easy to understand quickly, though it does not provide the same depth of information as BART.

Final Results:

In comparing T5 and BART for abstractive text summarization,various metrics were observed:

T5 demonstrated better performance in ROUGE metrics, with higher scores in ROUGE-1, ROUGE-2, and ROUGE-L, indicating better alignment with the reference content (which is an abstract 1 sentence summary of the news).
In BERTScore, T5 excelled in precision and F1 score, reflecting more semantically relevant and contextually aligned summaries, although BART had higher recall.

BART outperformed T5 in METEOR, suggesting better handling of lexical variations and semantic alignment. The comperatibly higher meteor score of BART could be due to the difference in their pretraining objectives. BART uses a more complexed span corruption that includes various span corruption techniques(sentence shuffling,document rotation, alongside token masking) .Could be making it better at word choices, order, and semantic variations comparing to T5.

Human evaluation technique demonstrated high differences:

Even though the models were initialized with the same max token number, BART's summaries were significantly lower and detailed comparing to T5's summaries, it even had a cut off. **BART's more detailed summarie**s, though informative, occasionally suffered from information overload. **T5, while concise and clear**, sometimes missed key details, affecting completeness.

BART excelled in detailed contextual information and lexical diversity. Also resulted in a higher meteor score which evaluates word choices, order, and semantic variations. These couldbe due to its denoising autoencoder architecture and pretraining objectives that include various span corruption techniques(sentence shuffling,document rotation, alongside token masking). These enhances its ability to understand and generate varied textual content.

Overall, **T5 is effective for [summarization] tasks requiring precise content alignment and concise summaries** this can be due to it's pretraining objective: text infilling where the model learns to predict the masked spans from the surrounding context.
And **BART is better for [summarization] tasks requiring detailed insights** could be due to it's pretraning objective denoising autoencoder. As it includes various span corruption techniques: sentence shuffling,document rotation, alongside token masking that could be making it better at word choices, order, and semantic variations comparing to T5 .

4.4 References

https://medium.com/analytics-vidhya/t5-a-detailed-explanation-a0ac9bc53e51

https://medium.com/dair-ai/papers-explained-09-bart-7f56138175bd

https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know