Evaluating the Efficacy of Pre-Trained Models and LSTMs in Abstractive Summarization of Reddit Posts

Steven Chang, Rebecca Sun
Data 266: Natural Language Processing
UC Berkeley School of Information
{stechang, rebecca.sun}@berkeley.edu

Abstract

The majority of text summarization tasks were primarily focused on formal text for training. However we expect that future of automatic text summarization to be in informal abstractive summarization. We explore the efficacy of pre-trained and finetuned models on informal text, in particular the tldr-17 Reddit dataset. Utilizing ROUGE and BERTScore, we see an improvement in style and tone of generated summaries, matching the original summaries.

1 Introduction

As the usage of machine learning expands, the need for a general automatic text summarization program will likely become increasingly reliant on the ability to generate abstractive summaries from informal text. As the vast majority of methods currently utilizes formal documents, the abstractive ability of these methods is diminished, due to inherent characteristics in formal text. Similarly, the amount of informal text heavily outweighs the amount of formal text, pushing us to explore abstractive informal text summarization. Given that prior efforts to consume informal text for summarization tasks have yielded much more abstractive summaries, we believe that the intake of these types of corpus will be the focus of automatic text summarizations.

As such we seek to explore the application of informal text in the field of abstractive text summarization, specifically in the context of social media and Reddit. In particular, our goal is to see the requirements needed to develop a model capable of producing automatic text summarizations of Reddit posts in the style of Reddit. Our original intention is to observe what is involved in developing an effective abstractive summarization model, with success being a definitive increase in the ROUGE score and BERTScore metrics in comparison to the reference summary.

2 Background/Literature Review

The larger focus of summarization usually utilizes more formal text summary pairs as the basis for training and evaluation, with corpora including CNN/Daily Mail, Newsroom, and XSum. Usually this allows for more structured and cleaner datasets, however this makes abstractive summarization a more difficult task as compared to extractive summarization, where summaries are done by extracting key sentences (Awasthi et al., 2021; Kim et al., 2019). Kim et al. finds that for abstract summaries of formal text, key sentences tend to be located at the beginning of the test, allowing summarization models to "simply memoriz[e] keywords or phrases from particular locations of the text." In fact, prior studies find that such abstractive methods trained on datasets with this characteristic tend to not show much abstraction (See et al., 2017).

In comparison, informal text provides much more abstract summaries and more uniform distribution of key sentences (Kim et al., 2019). Völske et al., 2017, who introduced the tldr-17 dataset, and Kim et al., who introduced the TIFU subreddit as a prelabeled data source, offered social media as a viable large corpus of informal text for summarization tasks. In both cases, the summaries were provided by the author of the original content, creating content-summary pairs where the summary (supposedly) contains the author's original intention. Völske et al., 2017 argues that pulling and creating this corpus from social media allows for a much more diverse genre for deep learning summarization. However they also find that these informal corpus require much more preprocessing before being usable. Indeed, in our own exploration of these datasets, we find that many human summaries fail to capture the context, or merely just contain the emotional sentiment of the content. Re-framing this issue positively, summarization for informal text is usually more abstractive as compared to for-

mal text (Syed et al., 2019).

With the focus on abstractive summarization on informal text, we should be aware of the faithfulness of the summary to the original text in both information and in fluency. This includes the vernacular and tone unique to the informal source, in this case Reddit posts. As such, we should consider the application of other metrics aside from ROUGE. Fischer et al., 2022 demonstrates a variety of metrics capable of fulfilling the role of automated faithfulness measurement, of particular note BERTScore and Entailment. The BERTScore is an evaluation metric measuring similarity between tokens utilizing the contextual embeddings (Zhang et al., 2020; Fischer et al., 2022).

3 Data Description

3.1 Data Source

The Webis TLDR-17 Corpus dataset was incrementally streamed and loaded, ensuring efficient management of its large size by retaining only crucial columns for summarization.

To prepare a balanced and specific subset of the Reddit dataset for training abstractive summarization models, we selected the 3 subreddits, "AskReddit", "worldnews", "funny". For each subreddit, it filters the dataset to include only posts from that subreddit with summary word counts greater than 10. By filtering and selecting posts based on subreddit and summary length, we aim to create a dataset that is both diverse (in terms of content from different subreddits) and relevant (with sufficiently summaries length for summarization tasks).

Our final dataset contains 36,000 records in total, with 12,000 from each of the specified subreddits, with another 1,500 records for testing.

3.2 Data Exploratory Analysis

By examining both the word cloud and the word count distribution plots, for ‘AskReddit‘, the largest words are "people," "like," and "one," suggesting discussions revolve around personal stories or opinions. The word cloud for the ‘funny‘ subreddit has prominent words like "make," "know," "say," and "go," implying a focus on storytelling. The statistic show shorter content and summary lengths, which fits with the punchy nature of humorous posts. In the ‘worldnews‘ word cloud, terms like "government," "country," "world," and "think" stand out, reflecting discussions on global events and politics. The stats confirm this, with mid-range content lengths and the longest summaries on average, indicating detailed discourse on complex topics.

4 Models

4.1 Initial Attempt

The lecture briefly mentioned the idea of the basic Encoder-Decoder recurrent neural network (RNN) architecture before delving into the pre-trained encoder-decoder transformer models such as T5. Our goal is to make a progressive approach to fully understand the “how” and “why”.

A traditional Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN)(Staudemeyer and Morris, 2019) reads sequences linearly, from start to finish, potentially missing future context cues. Contrastingly, a Bidirectional LSTM(Shah et al., 2021) examines data forward and backward, giving it a temporal panoramic view, in which equips the model with a richer context grasp, enhancing tasks like abstractive summarization, where understanding the intricate tapestry of text is crucial for crafting precise, comprehensive summaries. Bidirectional Long Short-Term Memory (BiLSTM) network, which is a type of recurrent neural network (RNN) architecture(Al-Sabahi et al., 2018), consists of two LSTM layers that process the data in both forward and backward directions to capture information from both past and future states. The inputs (x) are fed into both LSTM layers simultaneously, with the forward layer processing them from start to end, and the backward layer from end to start. The outputs of both layers at each time-step are then combined by concatenation, and passed through an activation layer to produce the final output sequence (y) (See Figure 1) .

Our model begins (See Figure 2) with input sequences that pass through an embedding layer to transform words into vector representations. A BiLSTM layer processes these embeddings, capturing dependencies and context from both directions of the text. The concatenated forward and backward states from this layer serve as the context for the decoder. (Chiu and Nichols, 2016) The decoder, also an LSTM but with a doubled latent dimension to accommodate bidirectional context, then generates a summary sequence. It starts with its own embedded inputs and iteratively predicts the next word, conditioned on the encoder’s context and its previous outputs. A dense layer with a softmax activation function assigns probabilities to each

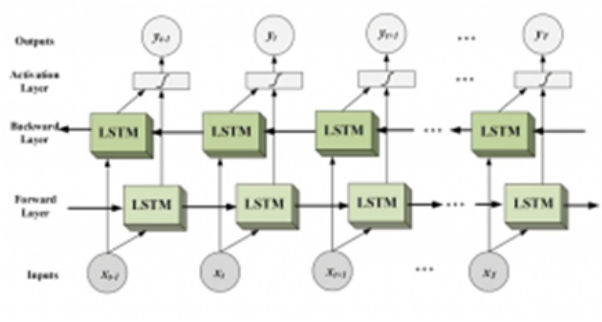

Figure 1: Structure of the Bidirectional Long Short-Term Memory (BiLSTM)

word in the vocabulary for the next word in the summary, resulting in a generative model that can abstractively summarize input texts.

4.2 Result and Error Analysis

The model's loss was 2.9656 during training and 3.1317 on the validation set, suggesting the model is performing slightly worse on unseen data. The ROUGE scores, which assess the quality of the text summarization, are quite low (around 5%), indicating the summaries generated by the model have limited overlap with the reference summaries. This suggests there's substantial room for improvement in the model's summarization capabilities.

We have observed that there's a noticeable discrepancy between the content and summaries in our training dataset, particularly regarding style, vocabulary, and context. Many of the summaries don't accurately reflect the meaning of the corresponding Reddit content. Given that ROUGE scores mainly evaluate the n-gram overlap between predicted and reference summaries, this inconsistency in content-summary pairing in our dataset suggests limited overlap. Consequently, this issue hinders our model's ability to grasp the text's core message and often results in irrelevant summaries. Additionally, the model faces challenges in maintaining the context and narrative flow of the original text, which becomes increasingly evident with longer and more intricate stories.

4.3 Pre-trained & Finetuned

As the initial attempt to build a model from scratch generated particularly poor sentences, we considered first training the model on a more higher quality dataset first, such as CNN or XSum. However, the resulting model would likely not be comparable in quality to already existing pre-trained models such as T5, Pegasus, and Bart. Similarly, training our initial model on another dataset before finetuning on tldr-17 would effectively be the same as finetuning on a pre-trained model.

Considering the limitations in resources for this project, we decided to instead set the new baseline to be these pre-trained models and finetuned on top of t5-small and bart-base. Finetuning on pegasus was also considered, however due to resource limitations, we decided not to. Finetuning was done using the ROUGE-1 metric for loss validation. The primary motivation was to have the model adopt the language and tone of these informal posts by encouraging the prediction of common and informal words.

5 Results

5.1 ROUGE and BERTScore

Human Summary	"Would dream about my current girlfriend years before I actually met her"
bart-base Pretrained	"Since I was about nine years old, I've been having a dream where I'm walking through the upstairs auditorium of my elementary school which is filled with kids I used to go to school with..."
bart-base Finetuned	"I had a dream where I met a girl I used to go to school with, and she was my girlfriend."

Table 1: Example of Summaries

Naturally, this improved the ROUGE metrics of finetuned models to their pre-trained counterparts (see Table 2). For example, comparing ROUGE-LSums of base-base pre-trained to finetuned, there's a 0.038 increase. However, one should consider the noticeable difference in summary tone (see Table 1). In general we noticed the pre-trained models tried to encapsulate the whole story of the content into its summary, while finetuned would more closely pick up on the key point of the content. Arguably, having a more complete picture of a content is what a summary should be achieving, but with informal social media texts, summaries are much more unstructured and informal, sometimes encapsulating more of an author intention than actual information.

In this respect, we might be able to leverage BERTScores as a metric for author intention and

faithfulness. We actually see that BERTScore improve for finetuned models as compared to pretrained (see Table 3). Assuming BERTScore, as a measurement on context embedding similarities, is an accurate measure of faithfulness to the author's intentions, as discussed by Fischer et al.[1], then we see that finetuning using ROUGE-1 has improved the faithfulness of the summary by a measurable amount. It could be argued that this increased probability of matching tokens is what lead to a better BERTScore, just by comparing similar embeddings. However, considering the baseline BERTScore were already fairly high while most generated summary doesn't fit the tone of the human summary, we should considering the finetuning on rouge to yield more natural informal text. We suspect that training on ROUGE has likely trained the model to adopt the summary's tone and vernacular, unique to internet speech, by adopting its common vocabulary.

5.2 Subreddits

Broadly, we have split the types of content based on the subreddit "funny," "AskReddit," and "worldnews," with the expectation that "funny" would be far more informal as compared to "worldnews," which should be more similar to formal text corpus. We expected to observe differences in changes in both ROUGE scores across the three splits between their pre-trained and finetuned counterparts, however ROUGE scores improved across the board (See Table 2). The only outlier was bart-base ROUGE-1 score for worldnews, which decreased between the pre-trained and finetuned. However all other metrics improved for the same model/split.

Based on BERTScores (See Table 3), we see that the F1 for "funny" and "AskReddit" splits experienced a greater increase as compared to "worldnews." If we consider that an informal summarization of a discussion on world news would closely resemble a formal summarization on the same subject that is in formal text corpus, then any improvement of the overall model on tldr-17 would see the most gain in the areas furthest from the training. For example, datasets like Newsroom may contain opinion articles and editorials with their own summarizations. As such these formal datasets may include pseudo-informal content-summary pairs; "worldnews" dataset may be similar to opinion articles and editorials, so finetuning on such examples would not yield as pronounced results as other informal subjects. A confirmation of this would require an indepth exploration of both tldr-17 and formal datasets, which is beyond the time and scope of this project.

5.3 Quality Issues

It should be noted that while BERTScores are relatively high across the board, the quality of the predictions are heavily limited by quality of the content-summary pairs. A cursory look through the dataset shows that while much of the tone and fluency of finetuned predictions match the overall tone of the human summaries, many predicted summaries are largely incongruous with the human summary. We find that within the original tldr-17 dataset, many summaries are of poor quality. This is noted by Völske et al., who also showed that proper data cleaning could result in a cleaner subset of examples. Unfortunately we were unable to replicate their work due to time and resource constraints, and instead filtered based on length of content and summary. This likely contributed significantly to overall quality of the predicted summaries.

6 Conclusion

We see that finetuning informal text corpus will improve the faithfulness of abstractive summaries to the original authors' intent. By finetuning on existing models using ROUGE metrics, we can train models to adopt the tone and style of informal text, in this case Reddit posts. This corresponds to a increase in BERTScore, showing that the faithfulness to the original author's intention can be achieved. There are issues however in the overall quality of the summaries generated, which we attribute to the poor quality of the dataset in general. Due to a lack of resources, we were unable to address this issue satisfactorily.

The quality would likely improve by cleaning the dataset and finetuning on a better subset of the tldr-17 dataset, as other studies have shown (Völske et al., 2017). At the same time, replicating the work by Zhang et al. on finetuning BERTScore would likely resolve to a more usable metric for measuring faithfulness, which could then be used as a metric for finetuning validation. However, it is unclear if this would improve on the style and

[1]Fischer et. al. utilized a finetuned version of BERTScore to get a correlation of faithfulness to more closely match human judgement. For our purposes, we are using the BERTScore evaluator provided by Huggingface.co

vocabulary generated in the summary. These next steps would likely help further the improvement of automatic informal text summaries.

References

Kamal Al-Sabahi, Zhang Zuping, and Yang Kang. 2018. Bidirectional attentional encoder-decoder model and bidirectional beam search for abstractive summarization.

Ishitva Awasthi, Kuntal Gupta, Prabjot Singh Bhogal, Sahejpreet Singh Anand, and Piyush Kumar Soni. 2021. Natural language processing (nlp) based text summarization - a survey. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pages 1310–1317.

Jason P. C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns.

Tim Fischer, Steffen Remus, and Chris Biemann. 2022. Measuring faithfulness of abstractive summaries. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 63–73, Potsdam, Germany. KONVENS 2022 Organizers.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks.

Sayed Rafay Bin Shah, Gavneet Singh Chadha, Andreas Schwung, and Steven X. Ding. 2021. A sequence-to-sequence approach for remaining useful lifetime estimation using attention-augmented bidirectional lstm. *Intelligent Systems with Applications*, 10-11:200049.

Ralf C. Staudemeyer and Eric Rothstein Morris. 2019. Understanding lstm – a tutorial into long short-term memory recurrent neural networks.

Shahbaz Syed, Michael Völske, Nedim Lipka, Benno Stein, Hinrich Schütze, and Martin Potthast. 2019. Towards summarization for social media - results of the TL;DR challenge. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 523–528, Tokyo, Japan. Association for Computational Linguistics.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

A Appendix

Following are attached tables and figures.

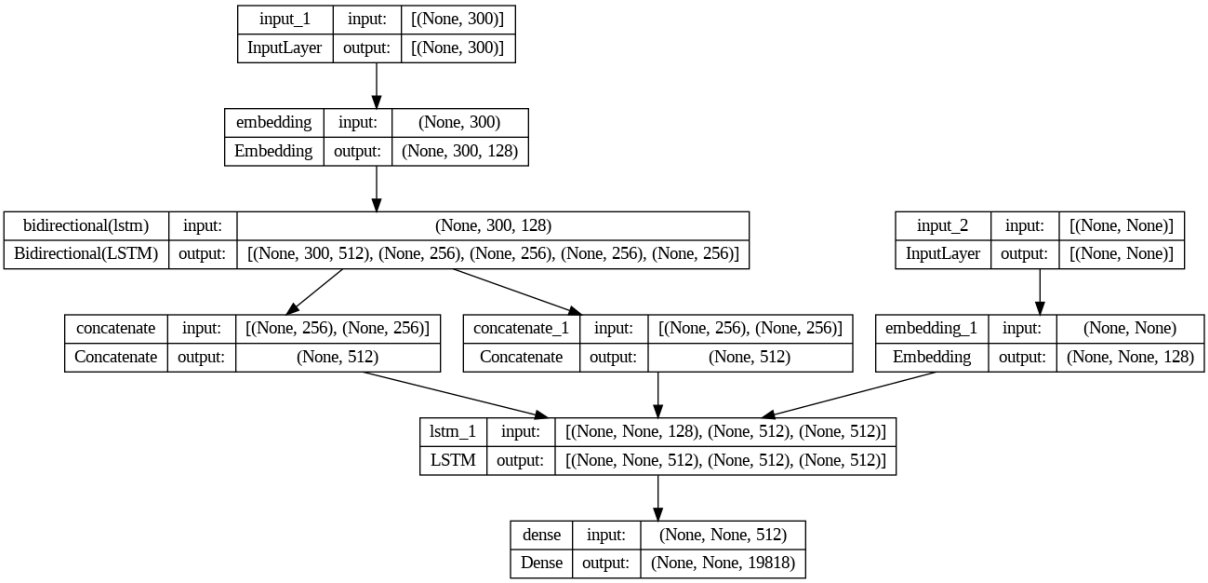

Figure 2: BiLSTM Model design

Model	Setting	Split	rouge1	rouge2	rougeL	rougeLsum
t5-small	Pre-trained	funny	0.17298	0.02544	0.11964	0.12198
t5-small		askreddit	0.19061	0.03750	0.13577	0.13804
t5-small		worldnews	0.18871	0.03310	0.12825	0.13190
t5-small		all	0.18418	0.03208	0.12777	0.13085
facebook/bart-base	Pre-trained	funny	0.16267	0.02910	0.10638	0.10937
facebook/bart-base		askreddit	0.15655	0.03294	0.10601	0.10850
facebook/bart-base		worldnews	0.18104	0.03240	0.11424	0.11934
facebook/bart-base		all	0.16678	0.03155	0.10879	**0.11247**
google/pegasus-xsum	Pre-trained	funny	0.12979	0.01834	0.10430	0.10507
google/pegasus-xsum		askreddit	0.13430	0.02191	0.10432	0.10549
google/pegasus-xsum		worldnews	0.14833	0.02253	0.11230	0.11522
google/pegasus-xsum		all	0.13757	0.02100	0.10681	0.10858
t5-small	Finetuned	funny	0.18728	0.03358	0.13691	0.13870
t5-small		askreddit	0.20620	0.04520	0.15642	0.15842
t5-small		worldnews	0.18631	0.03180	0.13146	0.13604
t5-small		all	0.19337	0.03676	0.14145	0.14431
facebook/bart-base	Finetuned	funny	0.17150	0.03824	0.13998	0.14149
facebook/bart-base		askreddit	0.21664	0.06052	0.17459	0.17577
facebook/bart-base		worldnews	0.16679	0.03720	0.13125	0.13387
facebook/bart-base		all	0.18520	0.04548	0.14849	**0.15057**

Table 2: ROUGE Scores of pre-trained and finetuned models on tldr-17 subsets

Model	Setting	Split	Precision	Recall	F1
t5-small	Pre-trained	funny	0.83070	0.84919	0.83973
t5-small		askreddit	0.83356	0.85558	0.84432
t5-small		worldnews	0.83340	0.84911	0.84106
t5-small		all	0.83256	0.85130	0.84171
facebook/bart-base	Pre-trained	funny	0.81801	0.85291	0.83496
facebook/bart-base		askreddit	0.81822	0.85597	0.83656
facebook/bart-base		worldnews	0.82008	0.85341	0.83627
facebook/bart-base		all	0.81877	0.85410	0.83593
google/pegasus-xsum	Pre-trained	funny	0.85155	0.84049	0.84571
google/pegasus-xsum		askreddit	0.85210	0.84331	0.84742
google/pegasus-xsum		worldnews	0.85360	0.84128	0.84718
google/pegasus-xsum		all	0.85241	0.84169	0.84677
t5-small	Finetuned	funny	0.85072	0.85124	0.85084
t5-small		askreddit	0.85526	0.85760	0.85632
t5-small		worldnews	0.85032	0.84898	0.84950
t5-small		all	0.85210	0.85261	0.85222
facebook/bart-base	Finetuned	funny	0.86439	0.84865	0.85627
facebook/bart-base		askreddit	0.87533	0.85830	0.86658
facebook/bart-base		worldnews	0.86436	0.84696	0.85540
facebook/bart-base		all	0.86802	0.85131	0.85941

Table 3: BERTScores of pre-trained and finetuned models on tldr-17 subsets