Objective-Oriented Transformer for Abstractive Document Summarization

Parma Nand[a], CangeGe Zhang and Manju Vallayil[b]
School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand

Keywords: Deep Learning, Language Modelling, Pre-Trained Transformer, Document Summarization, Generative Modelling.

Abstract: The effectiveness of transformer language models has been extensively used in a variety of language understanding and production tasks. Adaptation of these models for the specific purpose of text summarization has not been explored as much. In this work, we present the adaptation of a pre-trained Transformer model for the specific task of text summarization. A common way to train a language model is to randomly mask tokens from text and train the model to predict the masked words. The learner does this by paying attention to other neighbouring words in order to predict the masked words. Instead of training a single learner to learn random words, we trained three separate learners to focus only on specific types of words and generate separate summaries from multiple summary viewpoints. Then we used these focused learners to generate composite summaries corresponding to the type of words on which they were trained. We hypothesize that if we combine these different summaries, then it should result in a richer, more accurate summary covering multiple perspectives. We used already trained masked language models, BERT and RoBERTa, to extend the pretraining on the composite tasks of predicting just the nouns, the verbs or the rest of the words, as 3 separate pretraining objectives. We then trained the composite models for the downstream task of corresponding composite summarization. The evaluation was carried out by combining the three composite summaries with two benchmark data sets, Food Review and CNN/Daily Mail. The proposed composite pre-trained model and the composite summary generation algorithm produced a higher precision score based on ROUGE-1 and ROUGE-3 but a slightly lower score on ROUGE-2 compared to the state-of-the-art. The results showed that generating multiple summaries from different perspectives and then merging them has the potential to produce a richer and better summary compared to a one-shot strategy.

1 INTRODUCTION

During the pre-training phase of a Transformer (Vaswani et al., 2017), a Deep Learning Neural Network is presented with input/output sequence pairs and the target is to predict the output sequence, generated a word at a time, hence the name, generative model. The two sequences of texts can emulate the semantics between "The sky is cloudy and "It is going to rain. In this case, the Transformer will be emulating that cloudy sky results in rain, without explicitly specifying this semantic relation. Similarly, we can also train a Transformer to emulate the semantics among the words in the same sequence by presenting the same sequence with some of the words masked and forcing the Transformer to predict the masked words. This technique of training a transformer is referred to as a mask language model (MLM), (Devlin et al., 2018) and is self-supervised, as no human annotation effort is needed. This technique of training a transformer with huge amount of training data embeds generic text semantics within the neural network (NN) weights and hence can be used for a variety of text modelling tasks. However, we cannot expect the model to perform well on specific sequence-to-sequence (seq2seq) tasks, as it is trained only on very generic seq2seq data. For instance, it will not do well in generating answers from question texts as it has not seen these semantic relations in the training data. Hence, these types of generic models need to be further fine-tuned for specific tasks, known as downstream tasks, and all of the prior training therefore is referred to as pretraining. However, all these BERT-like models are designed for NLP tasks such

[a] https://orcid.org/0000-0001-6853-017X
[b] https://orcid.org/0000-0003-1962-8837

Nand, P., Zhang, C. and Vallayil, M.
Objective-Oriented Transformer for Abstractive Document Summarization.
DOI: 10.5220/0013682600004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 240-247

Table 1: Examples of composite summaries corresponding to the same source text.

Source Text: The first photographs have emerged of a stricken fishing vessel that has run aground on rocks and is spilling oil near an endangered penguin colony. The vessels owner says an investigation will need to uncover the "mistake that led to the accident. The 25m Austro Carina, owned and operated by Lyttelton-based Pegasus Fishing Ltd, ran aground near picturesque Shell Bay on the southeastern side of the Banks Peninsula on Sunday night. A helicopter recovered the skipper and three crew of the vessel, which was carrying 10,000L of diesel and 400L of hydraulic oil. Regional council Environment Canterbury (ECan) says initial aerial observations show oil from the vessel was headed towards Shell Bay and the neighbouring bays.
Summary 1: A fishing vessel, operated by Lyttelton-based Pegasus Fishing Ltd, has run aground near a penguin colony.
Summary 2: The 25m fishing vessel, Austro Carina ran aground spilling oil near Shell Bay.
Summary 3: A helicopter recovered the skipper and three crew from Austro Carina which ran aground.
Summary 4: A 25m fishing vessel has run aground and is spilling oil near endangered penguin colony.
Summary 5: A mistake has led to Austro Carina to run aground spilling oil which is drifting towards Shell and neighbouring bays.

as document classification, translation, text generation, and text understanding. To our knowledge, only one research group PEGASUS (Zhang et al., 2020) proposed a pre-training objective tailored solely for document summarization. The authors of this work proposed a self-supervised objective for pretraining a large Transformer-based encoder-decoder model on large text corpora. The aim of this task was to take out significant sentences from an input document and combine them into a single output sequence, similar to an extractive summary. It was discovered that masking entire sentences from a text and creating these gap sentences from the remainder of the document is an effective pre-training goal for downstream summarization tasks. The pre-trained model was tested on 12 different summarization tasks, including news, science, stories, instructions, emails, patents, and legislative bills. The results of the experiments showed that PEGASUS achieved the best performance in all 12 datasets, measured by ROUGE scores (Lin, 2004).

Working towards the aim of this multi-shot summary generation objective, we present three composite pre-training objectives for summarization and test the pre-trained model on the task of abstractive summarization task. The idea is inspired by the language learning pedagogy in which students are given an article to read together with a summary, with some of the key words blanked out. The learners are then asked to generate the summary by filling in the blanked words. To apply this concept to summarization, a semi-supervised dataset is used with variants of MLM by masking only nouns (MSLM-NOUN), masking only verbs (MSLM-VERB), and masking the rest of the words (MSLMREST). A further motivation for this strategy is based on the hypothesis that splitting attention to different syntactical elements will generate summaries on different perspectives. Instead of forcing a single learner to learn all aspects of the input, we split the task into three separate but related tasks to force them to learn each task better. A source text can give rise to multiple different summaries and all of them would have some degree of relevance as shown in Table 1. The summaries would all have different perspectives of the source text, depending on the intent of the reader, which could be more or less relevant. Forcing the learner to pick up critical nouns, verbs, or other tokens by paying attention to the appropriate tokens in the source text would generate the different summaries, focusing on entities (nouns), actions (verbs), and properties (adjectives, adverbs, and conjunctions). By splitting the tasks into specific objectives, we can better train individual learners to focus on critical parts of the source by paying extra attention to them. In this paper, we explore if forcing an already pre-trained learner to focus on different syntactical elements of the source text would generate summaries focused on the corresponding aspects of the source text, hence conflating the summaries at the end would give us a more "wholesome summary.

In summary, we propose a novel objective-oriented and pre-trained Transformer model for summarization based on the MSLM objective. We then evaluate the performance of this multi-shot model with the state-of-the-art one-shot models for summarization. We report that the F1 Rogue scores were higher for the multi-shot model compared to the state-of-the-art single-shot models.

2 RELATED WORK

2.1 Language Modelling

Language modelling (Wang and Cho, 2015) underpins tasks such as question answering, summarization, and translation. Most state-of-the-art models follow the seq2seq paradigm and are pre-trained on varied input-output sequences. Dai and Le (2015) introduced an RNN-based model trained with unlabelled data to learn next-word and sequence reconstruction, laying the foundation for downstream tasks like summarization.

Transformer-based models have since become dominant, using pre-training on large corpora and task-specific fine-tuning. For instance, MASS (Song et al., 2019) masks sentence fragments for reconstruction, while UniLM (Dong et al., 2019) combines masking, unidirectional, and seq2seq objectives. GPT-2 (Radford et al., 2019) focuses on next-token prediction, and BART (Lewis et al., 2019) corrupts and reconstructs input via shuffled segments, using [MASK] tokens for span annotations. Transformer-XL (Dai et al., 2019) improves context handling by encoding relative positional information, enabling longer, coherent generations.

While most abstractive summarization models rely on encoder-decoder architectures, Narayan et al. (2018) explored ConvNets for single-sentence summaries, an approach less evaluated in this context. Extensive studies, such as Rothe et al. (2020), have benchmarked transformer checkpoints for downstream tasks, finding that randomly initialized models may outperform BERTGPT2 combinations in some settings.

2.2 Transformer Pre-Training for Document Summarization

Document summarization (Yao et al., 2017) aims to produce concise versions of source texts. Early extractive methods treated it as a sentence ranking problem (Kupiec et al., 1995; Conroy and OLeary, 2001), selecting top-ranked sentences for the summary. SUMMARUNNER (Nallapati et al., 2017) was among the first to apply CNNs for this task. Recent advances such as PEGASUS (Zhang et al., 2020) and RoBERTa (Liu et al., 2019) established new state-of-the-art results using pre-trained transformers.

BottomUp (Gehrmann et al., 2018) blends word selection and abstractive generation, using a bidirectional LSTM for word salience prediction, incorporated into a transformer decoder. Generic models like BERT (Devlin et al., 2018) are also integrated into seq2seq summarization. TED (Yang et al., 2020), an unsupervised model leveraging the journalistic lead bias, was pretrained on 21.4M news articles using automatic filtering and denoising autoencoding. It outperformed other unsupervised baselines across three datasets.

2.3 Enhancing Summary Faithfulness

Pasunuru and Bansal (2018) improve abstractive summarization coherence via multitask reinforcement learning, combining entailment and summary generation tasks in a shared seq2seq model (Luong et al., 2015). Both tasks use a two-layer LSTM-RNN, with Gigaword (Graff and Cieri, 2007) as the summarization corpus and SNLI (Bowman et al., 2015) for entailment pairs. ROUGE evaluations show improved summary quality.

Cao et al. (2018) enhance factual accuracy by extracting open information tuples (Banko and Etzioni, 2007) from source sentences, converting them into fact descriptions. These are concatenated with the source text before summary generation. A dependency parser merges incomplete tuples to improve coverage. Generated summaries show a 40% higher usage of fact description terms.

We adopt ROUGE-N and ROUGE-L (Lin, 2004) to evaluate n-gram and longest common subsequence overlap between generated and reference summaries.

2.4 Masked Language Model (MLM)

The primary goal of RoBERTa pretraining is MLM, which is implemented in the same way as BERT but with slight variations. The model treats each sentence as an input and uniformly selects 15% of tokens from each sentence for replacement. From the tokens selected to replace, they randomly select 80% of the tokens to replace by using the special symbol [MASK], 10% original tokens are retained, and another 10% is replaced by randomly selected other “noise tokens. The target is to identify the original words that have been replaced. The goal of MLM is to predict the masked tokens by utilizing cross-entropy loss. In our project, we keep the token masking task similar to that of BERT and RoBERTa. A special symbol, [MASK], represented by the value 25,652 was added to the dictionary and 15% of the tokens were replaced with the mask symbol.

3 METHODOLOGY

3.1 Mask Summary Language Model (MSLM)

We introduce a new pre-training objective, named (MSLM), which is an extension of MLM designed to be better adapted for the downstream summary generation task. The MLM model is a generic model which uses a self-supervised objective, since it only needs the source text for masking and pretraining. The proposed MSLM, on the other hand, is specifically designed for downstream document summary tasks and therefore requires source and summary pairs as training data set. The pre-training phase aligns the gold standard summary text in addition to the source text with the target output. To do this, we kept the source text as original, but masked certain linguistic categories of words in the summary text. We then concatenated the masked summary to the source text to be used as input. The models objective is to predict the masked words in the summary, using only the unmasked words within the summary.

We designed three variants of MSLM based on three linguistic categories. MSLM-Noun, keeps all nouns as original and masks all other words; MSLM-Verb, keeps all verbs as original and masks all other words; MSLM-Rest, keeps all other words and masks all nouns and verbs. Note that the input sequence consists of the unmasked source text concatenated with the summary with either the nouns, verbs, or rest of the tokens masked. This resulted in three separate models focused on the three separate aspects of a summary.

3.2 Diverse Aspects Document Understanding (DADU)

DADU design is based on the common fine-tuning of a seq2seq model which makes use of the source text as the input and the summary text as the target; however, our target is not the whole original summary text. Instead, the target in our model is only certain types of words in the summary, while all the other words are masked. For example, in our experiment, we designed DADU-Noun, which means we only kept nouns as original in summary, using - instead of all verbs and using = instead of all the other words. The other two composite training models are illustrated in Figure 1. The Python Spacy library was used to identify nouns and verbs (POS types NOUN and VERB), while the remaining words were those that were not classified as NOUN or VERB.

The summaries generated for each of the three variations of the target represent three different perspectives of the source text, emphasizing the fact that different humans would also generate slightly different summaries for the same text, emphasizing different aspects of the information contained in the text. To obtain a complete summary, we then used a merge function (described in the next section) to obtain a complete summary of the input text.

3.3 Merge Function

The merge function will merge the output summary words into DADU-Noun summary by merging their corresponding symbols and the same order of the nouns, verbs, and rest of the words in the sequence. For example, the first symbol in DADU Noun is “=, which indicates that this position should be a word that has been classified in Rest sub-task. The first word in the DADU-Rest word list is This, therefore, the merge function will merge This into the = mask. In the case of a mismatch of the number of words generated versus the number of masks in the summary, the rest of the words or the remaining symbols were discarded. Similarly, the DADU- Verb is used to complete the noun summary. The same procedure was followed to complete the summaries of DADU-Verb and DADU-Rest, illustrated in Figures 2 and 3.

4 RESULTS AND ANALYSIS

4.1 Datasets

We used two publicly available datasets: Amazon Food Reviews and CNN/Daily-mail datasets, briefly described below. The Amazon Food Reviews dataset consists of food reviews from Amazon. It spans a period of more than 10 years (October 1999 - October 2012), consisting of a total of 568,454 datapoints. The reviews include product and user information, ratings, and a plain text review per datapoint. There is also a summary of the reviews which can be used as the gold standard for abstractive summary. The CNN/Daily Mail dataset contains 93k CNN articles and 220k Daily Mail stories, and their corresponding human-generated abstractive summaries as bullets. In total, the corpus has 286,817 training pairs, 13,368 validation pairs, and 11,487 test pairs, as defined by their scripts. We split the data into two equal parts, one for pretraining and another for fine-tuning, each part containing almost the same number of data points.

DADU-noun target summary: Keep all nouns as original, use '-' instead all verbs, use '=' instead of all the others.		= = Secretary = State - = = = decision = - = Navy SEALs = - Osama bin Laden. - = she - = it = Biden = = =. = she = Biden = - = = runners = = race.
DADU-verb target summary: Keep all verbs as original, use '+' instead of nouns, use '=' instead of all the others.		= = + = + spoke = = = + = order = + + = kill + + +. Said = + pushed = + = + = = =. = + = + = seen = = + = = +.
DADU-rest target summary: Keep all other words as original, use '-' instead all verbs, use '+' instead of nouns.		The former + of + - out about the + to - the + + to - + + +. - that + - for + and + was more cautious. Both + and + are - as front + in 2016 +.

Figure 1: DADU input and target for the three sub-tasks.

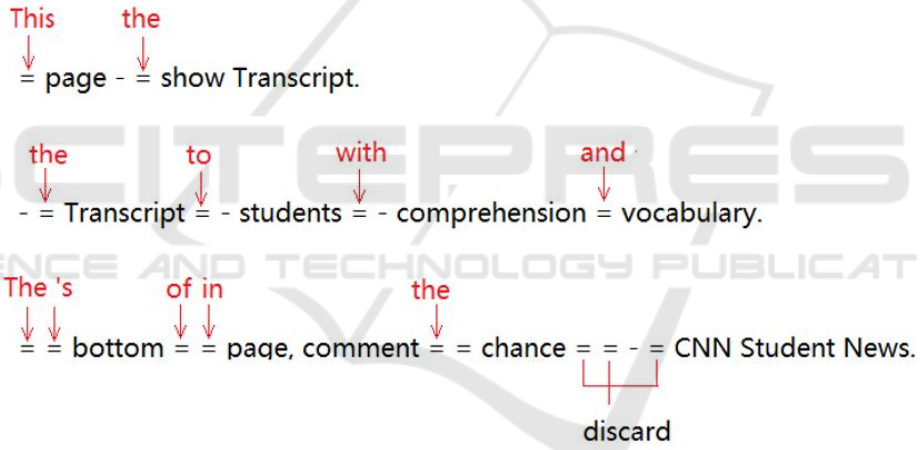

Figure 2: Merge Rest into DADU-Noun summary.

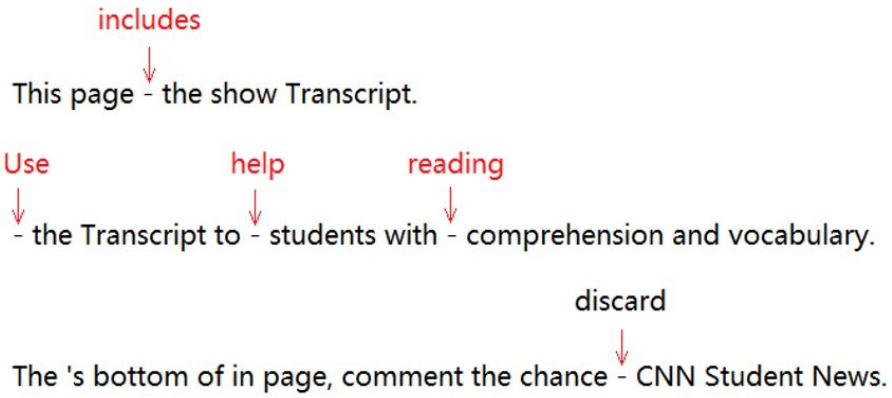

Figure 3: Merge Verb into DADU-Noun summary.

Table 2: The details of model comparisons.

Model	Objective	Dataset	Trained Epochs
BERT	MLM + NSP	Wikipedia	40 epochs
RoBERTa	MLM	Wikipedia	40+ epochs
RoBERTa-CNN20epochs	MLM	CNN/Daily-mail	20 epochs
BQ	MLM + MSLM	CNN/Daily-mail	20 epochs
BQ-partial-Noun	MLM + MSLM-Noun	CNN/Daily-mail	20 epochs
BQ-partial-Verb	MLM + MSLM-Verb	CNN/Daily-mail	20 epochs
BQ-partial-Rest	MLM + MSLM-Rest	CNN/Daily-mail	20 epochs

4.2 Data Pre-Processing

Since we used RoBERTa as our benchmark comparison, we filter text larger than 500 tokens, since RoBRTa has a limit of 512 tokens. We used the Food Review dataset only for the downstream Abstractive document summarization task and used the CNN/Dail-Mail dataset for both pretraining and summarization tasks.

4.3 Model Training

All our experiments were trained on Google Colab; however, the computing resource provided by Colab is limited and sometimes unstable, thus, we restricted our pretraining tasks to 20 epochs, and finetuning tasks to 10 epochs for each experiment. Since the training time for our experiments has been restricted compared to some of the state-of-the-art works (e.g., RoBERTa was trained for over 40 epochs), the final evaluation values reflect this. Hence, to compare the results on the same dataset and training time, we pretrained RoBERTa on the same CNN/Dail-Mail dataset (the original was trained on Wikipedia). We named this version RoBRTTa as RoBERTa-CNN-20 epochs. Table 2 outlines the various names and the training epochs for BERT, RoBERTa, and BQ (our model, abbreviated for "Blank Quiz" which forms the motivation for composite task oriented approach) models. Note that, as a comparison, we also trained our BQ model with random masking of words, as is the case with standard masked language modelling. The base used for all BQ variants in our modelling was the RoBERTa model.

4.4 Results for Search of the Best BQ-Partial Variants

As described earlier, we first pre-trained each BQ-partial variant on CNN/Daily-mail dataset for 20 epochs. Then, we fine-tuned them for summary generation on the Food Review and CNN/Daily- mail datasets for 10 epochs each, respectively. The fine-tuning was tested by using each of the variants to generate the complete summary for the two datasets. The results shown in Table 3 show that the best performer among the variants was the BQ-partial-noun.

Table 3: Comparison of the various BQ variants.

	BQ-partialnoun	BQ-partial-verb	BQ-partialrest
Food Review-fine-tune			
Rouge-1	18.22	15.17	14.79
Rouge-2	5.28	4.11	3.83
Rouge-L	18.02	14.99	14.65
CNN/Daily-mail-finetune			
Rouge-1	24.95	5.64	3.27
Rouge-2	0.77	0.21	0.12
Rouge-L	23.63	5.31	3.07

Table 4: Comparison of BQ and RoBERTa trained for 20 epochs.

Model	**Rouge-1**	**Rouge-2**	**Rouge-L**
RoBERTa-CNN20epochs	27.53	8.41	19.68
BQ	30.18	9.67	21.35
BERT	42.13	19.60	39.18
RoBERTa-base	42.30	19.29	39.54
RoBERTa-large	43.06	19.70	40.16

This reaffirms the intuition that nouns are the highest carriers of information among the various linguistic components of a language. Since the BQpartial-noun was trained on the MSLM-noun objective and delivered the best results, we chose to concatenate MLM and MSLM-noun to train our final BQ model. This is intuitive, as nouns usually have high information content that indicates the entities in the text. The verbs and all the rest of the words, such as conjunctions and adjectives, usually indicate the relationships between the entities. The findings demonstrate that nouns in texts are essential components for building models for common language tasks.

4.5 RoBERTa-CNN-20epochs vs BQ

To directly compare the effectiveness of the partial variants (Noun), we trained the BQ model with the MLM + MSLM noun objective for 20 epochs and compared it with RoBERTa-cnn-20 epochs on the MLM objective. The results are tabulated in Table 4. We have also included the numbers for the stateof-the-art models trained for a higher number of epochs in the bottom three rows for relative comparison. From the first two rows, BQ has consistently performed better compared to the equally trained RoBERTa version, RoBERTa-CNN-20epochs, which was the base for our BQ model. This is a promising result that indicates that with enough training, the model has the potential to perform better than the current state of the model, as the performance of the BQ model is approximately 75% of the state-of-the-art with less than half the training.

4.6 Summarization Comparison

In order to gauge the effectiveness of the proposed DADU model we trained the RoBERTa-based model on CNN/Daily-mail dataset for ten epochs, then we

Table 5: Results for DADU with one-shot summary generation.

	Precision	Recall	F-measure
Rouge-1 One-Shot (RoBERTa-base)	33.15	42.46	36.02
Rouge-2	15.95	18.97	16.58
Rouge-L	24.89	31.44	26.77
Rouge-1 DADU (Composite RoBERTa-base)	37.10	28.22	31.10
Rouge-2	0.85	6.38	7.04
Rouge-L	25.20	19.31	21.20

used it to generate the summaries on the CNN/Daily-mail dataset with the existing one-shot summary generation and the proposed composite summary generation model, DADU. The results are tabulated in Table 5. For comparison, we also finetuned a traditional seq2seq model based on RoBERTa-base with CNN/Daily-mail dataset for 10 epochs.

DADU obtained an approximately 4% improvement in the precision value of Rouge-1 and about 0.3% improvement in Rouge -L. However, it got a much lower score of Rouge-2 and all recall values, leading to the lower overall F1 score. In this project, our focus was to explore whether the generation of perspective-based composite summaries could assist us in generating more accurate summaries, and hence we used a very crude merging function to merge the three composite summaries. For merging, we essentially substituted the words based on the number of tokens in the gold standard, without any consideration for grammar or coherence. This means that although we managed to get the individual words correct, we could not generate grammatically correct sentences, or n-grams, from them, resulting in low ROGUE-2 and ROGUE-L scores. We also discarded the extra words that did not fit the generated sequence, which would have resulted in low recall values in Table 5. These preliminary results from the DADU tests show that the strategy could be powerful if integrated with a better merge module that is able to generate grammatically correct sentences given a set of words, hence also rearranging words giving better n-gram alignment. Table 6 shows some examples of label summaries and corresponding generated summaries for comparison.

5 CONCLUSION AND FUTURE WORK

We proposed a new objective for pre-training a transformer model on the masked language model. We tested the performance of the three variants of proposed objectives and found that one of them, MSM-Noun, gave a better comparative performance on the state-of-the-art. Based on the pre-training objective, we proposed a new fine-tuning method DADU for document summary generation. We showed that its performance was also higher for ROGUE-1, but slightly lower for longer sequences. The results showed that this was due to a very crude merge function which merges the words to construct the final summary, based on order and sequence, hence frequently resulting in incorrect n-grams even though the correct words might be present in the whole output. Both of the newly proposed models have shown encouraging results and need further rigorous testing on more powerful computing resources to surpass the state-of-the-art. As part of further work, we are working on integrating a better Natural Language Generation (NLG) based merge function so that this will be able to construct grammatically correct sentences from the words generated from the summarization phase. This will substantially improve the ROGUE scores as there will be more accurate matching of the n-grams with the reference text.

REFERENCES

Banko, M. and Etzioni, O. (2007). Strategies for lifelong knowledge extraction from the web. In *Proceedings of the International Conference on Knowledge Capture (K-CAP)*, pages 95–102.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642. Association for Computational Linguistics.

Cao, Z., Wei, F., Li, W., and Li, S. (2018). Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Conroy, J. M. and OLeary, D. P. (2001). Text summarization via hidden markov models. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 406–407.

Dai, A. M. and Le, Q. V. (2015). Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, volume 28, pages 3079–3087.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H.-W. (2019). Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32.

Table 6: Examples of DADU-generated and labeled summaries.

Label Summary	Generated Summary
Accident happens in Santa Ynez, California, near where Crosby lives. The jogger suffered multiple fractures; his injuries are not believed to be life-threatening.	A driving accident occurred in Santa Ynez, California. Crosby struck a jogger who suffered multiple fractures and was airlifted to a hospital in Santa Barbara.
Sigma Alpha Epsilon is being tossed out by the University of Oklahoma. Its also run afoul of officials at Yale, Stanford, and Johns Hopkins in recent months.	Sigma Alpha Epsilon is under fire for a video showing fraternity members chanting a racist phrase. The party that the video was made at has been suspended and the fraternity is facing permanent ban. This controversy is the latest in a series of incidents involving SAE in recent months, with hard work being done by alumni and chapters to make changes.
Join Meerkat founder Ben Rubin for a live chat at 2 p.m. ET Wednesday. Follow @benrbn and @lauriesegallCNN on Meerkat. Use hashtag #CNNInstantStartups to join the conversation on Twitter.	Meerkat launched during SXSW in Austin last week. Ben Rubin, the founder of Meerkat, will be available for questions and comments on Meerkat or Twitter. This is a great opportunity to ask him anything you want to know about the app.

Gehrmann, S., Deng, Y., and Rush, A. M. (2018). Bottom-up abstractive summarization.

Graff, D. and Cieri, C. (2007). English gigaword, third edition. https://catalog.ldc.upenn.edu/LDC2007T07. LDC2007T07.

Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pre-training approach.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2015). Multi-task sequence to sequence learning.

Nallapati, R., Zhai, F., and Zhou, B. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Ranking sentences for extractive summarization with reinforcement learning.

Pasunuru, R. and Bansal, M. (2018). Multi-reward reinforced summarization with saliency and entailment.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Rothe, S., Narayan, S., and Severyn, A. (2020). Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2019). Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning (ICML)*, pages 5926–5936.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, ., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Wang, T. and Cho, K. (2015). Larger-context language modelling. *arXiv preprint arXiv:1511.03729*.

Yang, Z., Zhu, C., Gmyr, R., Zeng, M., Huang, X., and Darve, E. (2020). Ted: A pretrained unsupervised summarization model with theme modeling and denoising.

Yao, J.-g., Wan, X., and Xiao, J. (2017). Recent advances in document summarization. *Knowledge and Information Systems*, 53(2):297–336.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning (ICML)*, pages 11328–11339.