Multi-task Transfer Learning for Sentimentally consistent Summarization

Sai Ashish Somayajula
University of California, San Diego
A59002565
ssomayaj@ucsd.edu

Arth Dharaskar
University of California, San Diego
A59004566
adharask@ucsd.edu

Abstract

This project aims to build a language model that produces summaries of articles that are sentimentally consistent with the original article. We introduce a novel loss function to encourage the model towards a more consistent summarization space. We evaluated various candidates for this loss function and settled for Mean Squared Error and KL-Divergence, receiving a BLEU score of ***0.16*** *and* ***0.17*** *respectively on the CNN Dailymail Dataset. We provide methods and a discussion on the various experiments in the respective sections.*

1. Introduction

In today's fast-paced life, we often want to consume as much content as possible, but in doing so, we often forget that every data point we consume affects our emotional state [3], [12]. The current state-of-the-art text summarization techniques [11], [9], [1] pay no heed to the original article's Sentiment while constructing the summary. They may not provide the same emotional impact as the original article. In this project, we tackle this problem by introducing a novel Sentiment Loss which constraints the summary space to align with the original article. We propose a multitask transfer learning-based model which is trained in two phases; phase one consists of pre-training a classifier network (Sentiment Head) for sentiment classification, and phase two consists of fine-tuning a BART based model(Summarization Head) which has been pretrained on a sentence-denoising task by Facebook [5]. The Sentiment Head's pre-training is performed on the IMDB dataset [6] while Summarization Head's fine-tuning is performed on the CNN DailyMail dataset [10].

We hypothesize that by introducing a sentiment consistency network and training the network in a multitask fashion, we will be inducing a sentiment prior over the summarization space. The final model is capable of producing not only a summary that is aligned with the original article but also a sentiment score for both the summary and the original article, which can be used for further downstream tasks such as news/social media recommendation, brand monitoring, customer service, and market research.

Our project has the following contributions:
1.We introduce a novel Sentiment Consistency loss.
2.We run various experiments with different Sentiment Consistency loss and provide ablations.
3.We develop a model that provides a sentimentally consistent summary and sentiment scores for both article and summary, which can be used for further downstream tasks.

2. Related Works

Text summarization is a problem that has been extensively studied. A **summary** [9] is defined as a text that can consist of one or more texts that convey the vital information of the original article and is no longer than half of the original article and is usually significantly less. Various traditional methods are consisting of Sentence scoring based on word frequency [11], using sentence embeddings [2], and skip-thought vectors [10]. However, considering the latest deep learning boom, methods are using attention-based sequence to sequence models [13], RNN, and gating-based architectures [4] and using reinforcement learning for abstractive Summarization.

The second task that is addressed is that of **Sentiment** classification which consists of predicting a score/scale that represents the tone, positivity, or polarity of an article. There have been various methods involving SVMs, LDA, Bag of words modelling, and Markov blanket-based methods [6], [7] to perform sentiment classification. However, current methods incorporate language models with a fine-tuned head at the top to perform classification.

Multitask learning(MTL) is an upcoming subfield within machine learning that simultaneously trains multiple tasks. The effect of training on a multitask objective reduces overfitting as the model tries to minimize loss with respect to one or more losses in the same shared parameter space. Training in a multitask setting also can be thought of as inducing a prior with other loss functions on the primary loss function. There are also considerable speedups in training in multitask settings if the auxiliary tasks support the primary task [15].

Transfer Learning is the process of using a pre-trained model and adapting its domain for another task. This is common in scenarios where there is not much-labeled data available, or the pretrained task is sharing a common representation as the target task. Transfer learning is ubiquitous in NLP(Natural Langauge Processing) as there are common linguistic representations and structural similarities in a language. Since the pretrained model already contains a language structure, the model's training proceeds at a faster pace as opposed to learning a representation from scratch, and it may also converge to a better local optima [14].

Our model draws inspiration from the current advances in deep masked language modelling architectures-Transformers that use multi-headed attention [12]. We attach a fine-tuned classification head at the top to provide the sentiment scores. We did not find any model that tackled both these tasks in a multitask fashion during our research. We will be describing the model architecture and method of training in section 3, the experiments we ran in section 4, results in section 5, and conclusion in section 6.

3. Methods

3.1. Dataset

3.1.1 IMDB

We use the IMDB dataset [6] for training the sentiment head on the sentiment classification task. It is a movie review dataset, partitioned into positive and negative movie reviews, consisting of a 25K train and 25K test dataset. We use the IMDB dataset from the Huggingface-Datasets library.

3.1.2 CNN DailyMail

We use the CNN Daily Mail dataset [10], which consists of 300K original articles and their associated summaries, for fine-tuning the BART model to produce the sentimentally consistent summaries of articles.

3.2. Model Architecture

This section describes the model architecture in detail.

3.2.1 Conditional Text Generation model

We use the BART model by Facebook as the Conditional text generation model, which generates summaries from the original input article. The BART mode [5] is a denoising autoencoder that aims to perform the task of sequence to sequence generation. It is trained by corrupting the input sentence and learning to reconstruct the original sentence. We use a summarization head on top of the decoder for the summarization task. It can be generalized as having a BERT model as an encoder and a GPT model as a decoder for generating the target texts. With the advent of the BERT model, it is possible to transfer learning in Natural Language Processing. We fine-tune the pre-trained BART model from Facebook on our dataset.

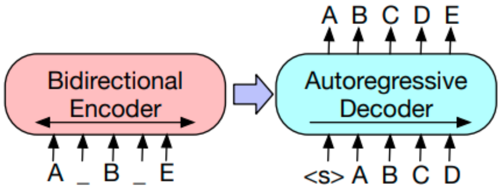


Figure 1. BART Model Architecture [5]

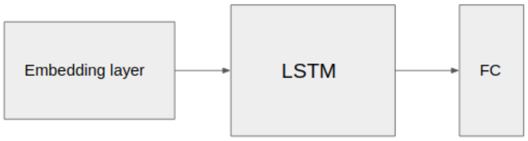


Figure 2. Classifier Model

3.2.2 Classification model - Sentiment Head

This subsection defines the classification model's architecture for defining the sentiment score and its associated loss—the classifier's architecture comprises of embedding layer, an LSTM layer, and a fully connected network. The LSTM layer consists of dropout(=0.5) and weight norm layers to prevent overfitting. We train the classifier on a sentiment classification dataset such as the IMDB dataset, discussed further in section 4. We then use this trained model on the sentiment dataset for defining the sentiment scores and associated losses.

3.2.3 Merged Model for Sentimentally consistent Summarization

We have the BART model as a conditional text generation model that produces summaries from the input articles. We have the classification model that acts as a sentiment head on top of the BART model that predicts the input and its corresponding summary sentiment scores. Figure 3 shows the merged model. We pass the input article through the BART model to obtain its summary. We then pass the input article and its summary through the sentiment head to obtain the sentiment scores. Given these sentiment scores on the input article and the generated summary, we define a sentiment loss intending to minimize the distance between them while training the BART model. This distance can be

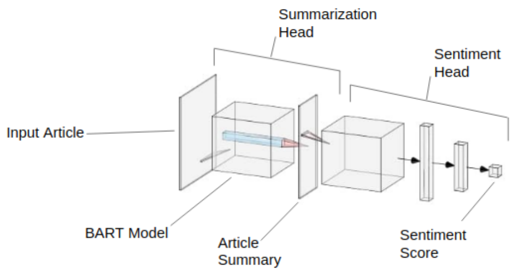


Figure 3. Merged Model

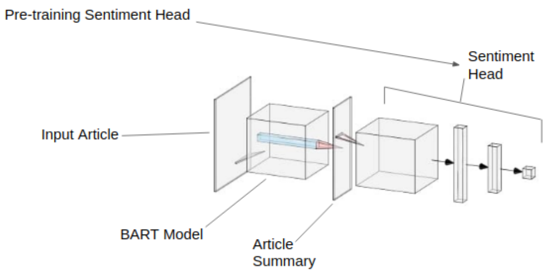


Figure 4. Stage-1 Learning

viewed as a distance in the euclidean space or probability distribution sense. Hence, we define two types of losses on the sentiment scores to account for both the views of distance measure and evaluate their performances in 4. The details of the sentiment loss function are given in the section 3.3.

3.3. Losses

This section defines the losses used to fine-tune the BART model to output ”Sentimentally” consistent summaries of the input articles. We effectively have two types of losses in this work.

3.3.1 Cross-Entropy loss

The loss is used to fine-tune the BART model and train the classifier model on their respective datasets. It is defined between the output logits $\hat{y}$ from the model and the target class y.

$$CCE(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{i} \mathbf{y}_i log(\hat{\mathbf{y}}_i) \tag{1}$$

3.3.2 Sentiment Loss

As stated in 3.2.3, after we obtain the sentiment scores from the merged model corresponding to the input article and the summary, we define the sentiment loss over the pair of sentiment scores. We use two different approaches while framing the sentiment loss.

3.3.2.1 MSE-Sentiment Loss

After obtaining the article and the summary sentiment scores, we view them as vectors in $\mathbb{R}^2$. The goal now is to minimize the euclidean distance between them. Hence, we define the sentiment loss as the MSE between the pair of sentiment scores.

$$MSE(\hat{y}_i, y_i) = \|y_i - \hat{y}_i\|^2$$

3.3.2.2 KL Divergence-Sentiment Loss

We view the sentiment scores as probability vectors over the classes (positive and negative). We require the target summary probability vector to be ”sentimentally” as close as possible to the actual probability vector corresponding to the article.

$$D_{KL}(p, q) = \sum_{x} p(x) \mathrm{log} \frac{p(x)}{q(x)}$$

3.4. Training Procedure

This section describes the steps in the training procedure. We decided to train the model in multiple ways, two of which are outlined below.

3.4.1 Method 1.

The training of the model is carried out in a two-folded fashion.

3.4.1.1 Pre-training the Sentiment Head

We first train the sentiment head on the IMDB dataset in a fully supervised fashion with the losses mentioned in Section 4.2. We can see that in Fig 4.

3.4.1.2 Training the BART model

In this stage of learning, we fine-tune the BART model on the Text Summarization Dataset. We train the model on the sentiment loss on the input article and summary’s sentiment scores. As for the sentiment head, we fine-tune the sentiment head for a few epochs on the sentiment dataset and also on the sentiment loss. The reason is that the Text Summarization Dataset is sentimentally consistent with the original articles. We also have the loss on IMDB dataset to ensure it doesnot overfit on the sentiment loss. Fig 5 shows the training of this stage.

After a certain number of epochs, we freeze the sentiment head weights and further fine-tune the pre-trained

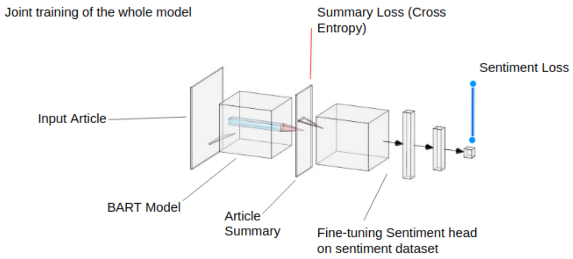

Joint training of the whole model
Summary Loss (Cross Entropy)
Sentiment Loss
Input Article
BART Model
Article Summary
Fine-tuning Sentiment head on sentiment dataset

Figure 5. Stage-2(a) Joint Learning

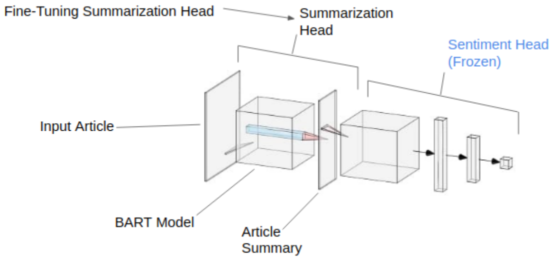


Figure 6. Stage-2(b) BART Model Learning

BART model on the Text Summarization Dataset and the sentiment loss. This can be seen in Fig 6.

3.4.2 Method 2.

We train the entire model together but it is observed that it takes a long time to converge. Details are explained below, 3.4.2.1.

3.4.2.1 Joint training the model

We pose the problem of joint training in which we have multiple optimizers trained together. The optimizers are updated individually on their respective datasets. The benefit this provides is that the Classifier head is tuned along with the BART model and provides a standard representation for both models. However, it takes longer to train using this method.

3.4.3 Evaluation Metrics

A common metric is the Bilingual Evaluation Understudy Score (BLEU-Score) to evaluate the computer-generated natural language. This score ranges from 0 to 1 and indicates a perfect match by '1' and a perfect mismatch by '0'. It has been developed initially to compare the Text's translations but has many different use cases in natural language processing, including evaluating generated captions [8]. The BLEU score is based on a concept called n-grams. An n-gram denotes a sequence of n subsequent words. The BLEU score uses those n-grams to describe a modified version of precision to compare different sentences. The precision is modified in the sense that the number of appearances of an n-gram in the reference sentence set determines how much credit a word can get in the final output prediction y. This procedure is denoted by $count_{clip}$.

$$p_n = \frac{\sum_{ngram \in y} count_{clip}(ngram)}{\sum_{ngram \in y} count(ngram)} \tag{2}$$

Where p_n denotes the BLEU score on n-grams only, the overall BLUE score is the exponentiated weighted average for short prediction lengths.

$$BLEU - N = BP * exp(\frac{1}{N}\sum_{n=1}^{N} p_n) \tag{3}$$

where BP is the penalization term defined by

$$BP = \left\{ \begin{array}{ll} 1 & \text{if len(y)} > \text{len(t)} \\ exp(1 - \frac{len(pred)}{len(target)}) & \text{else} \end{array} \right\}$$

where $len(y)$ and $len(t)$ denote the sentence length of the prediction and the target respectively.

We also check the absolute difference between the generated summary and the original article's Sentiment (Mean Absolute Sentiment Difference).

$$MASD = |ss(\text{Original article}) - ss(\text{Generated caption})| \tag{4}$$

Where ss is the Sentiment Score provided by the classifier head.

4. Experiments

To test our training methods' efficacy, we evaluate our models on the BLEU score and MASD as mentioned in 3.4.3, our models are trained using a train test split of 12k training examples and a 1k testing examples. The max sequence length for encoding the was chosen to be 500 for the article and 100 for summary. For the MSE loss we evaluated two different weightings of cost function while for the KL Divergence since the loss fluctuated a lot, any weighting that gave more weight to KL Divergence caused the exploding gradients problem. Weighted MSE uses a weight of 2 for the cross-entropy loss while a weight of 1 for MSE loss, while the balanced MSE uses weight of 1 for each of the losses. Weighted KL uses a weight of 1 for cross-entropy loss and a weight of 1e-2 for the KL divergence. The optimizer used for training the summarization head was SGD

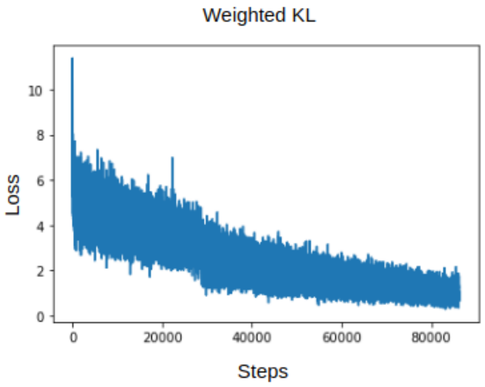


Figure 7. Weighted KL Divergence training loss

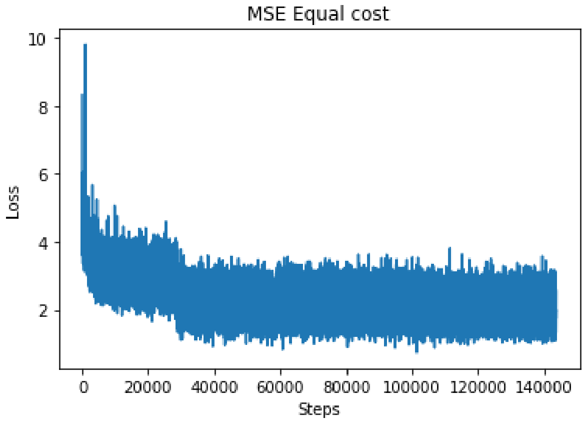


Figure 8. Equal weight MSE training loss

with a momentum of 0.9 while for training the sentiment head ADAM optimizer was used. We trained using a batch size of 4-6 for 50 epochs which took 12-14hours on a GTX 1080. We show the loss curves for the above mentioned models 7, 8, 9.

5. Results

In this section we discuss the results of our methods. Our hypothesis that training in a multitask setting holds true as shown in Table 1. We trained a baseline model with the same settings as the other models but without the sentiment consistency loss and we can see that this model performs worse both in terms of BLEU score and MASD. The gap between our best model-weighted KL divergence and the baseline model without sentiment score is of 47% also, the BLEU score is better by 13%. Within our sentimentally consistent models, KL divergence based Sentiment loss seemed to perform the best, we feel that it may be a result of the summarization network trying to make

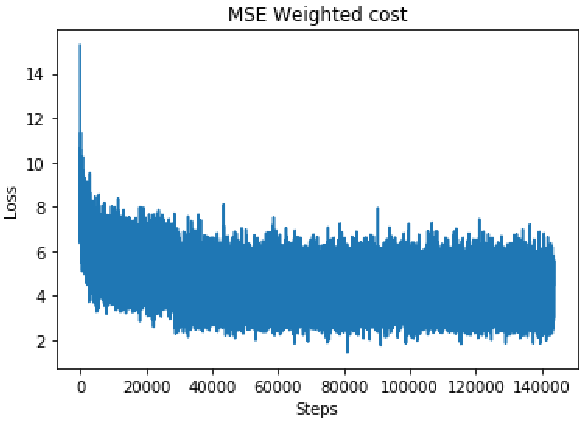


Figure 9. Weighted MSE training loss

the distribution over the sentiment scores as close as possible to the target sentiment distribution that may have lead to this. We also noticed that it may be the case that since MSE is more sensitive to outliers that the gradient signal can be skewed by such data points. In order to counter this while still considering the MSE signal if we do give more weight to cross-entropy summarization loss, we see similar BLEU scores for both cases but drastic improvement in the MASD. We notice from the results in Figure 10/Figure 11. that the model has learnt to extract certain features of the original articles that are not present in the ground truth. For eg, the first article in Figure 10, The ground truth contains no mention of the name of the brother but our model is able to infer that the brothers name is Anthony and includes it in the summary.

Model	BLEU	MASD	Loss
Baseline model	0.15472	1.65803	1.5689
Weighted MSE	0.16021	0.04265	3.0405
Balanced MSE	0.16022	0.21982	1.7832
Weighted KL	0.17555	0.03472	0.9781

Table 1. Testing Results

where, BLEU-Higher is Better and MASD-Lower is better.

Our model is however not without fault, we show some of the failure cases as well in Figure 10. and Figure 11. We see that the model sometimes exhibits poor sentence structure and ends up repeating a lot of words in order to fill the padding. We believe that these issues can be alleviated with regularization and training for longer with a bigger subset of dataset.

We also note that while BLEU score is the standard that is used in the NLP community it might not be suitable for evaluation on our task since it is typically used in language translation and not knowledge summarization. We did not find a metric that characterized the information density provided by a given text hence we report on the BLEU score.

6. Conclusion

In this project we successfully showed that training a language model in a multitask setting to generate summaries which were sentimentally consistent was plausible. We evaluated different loss functions and gave empirical evidence that KL divergence as a sentiment loss function performed the best. We also introduced a new metric MASD (Mean absolute Sentiment Difference) which characterized the difference in sentiment of article and generated summary and demonstrated that models trained with sentiment loss performed drastically better than those without. Future work could include a parameterization of weights for the losses and experimenting with different architectures for the classifier. We are also thinking of ways to characterize the information content present in a given text which could be used to note how much information loss takes place after a particular summarization.

References

[1] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*, 2017.

[2] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[3] Wendy M Johnston and Graham CL Davey. The psychological impact of negative tv news bulletins: The catastrophizing of personal worries. *British Journal of Psychology*, 88(1):85–91, 1997.

[4] Minsoo Kim, Moirangthem Dennis Singh, and Minho Lee. Towards abstraction from extraction: Multiple timescale gated recurrent unit for summarization. *arXiv preprint arXiv:1607.00718*, 2016.

[5] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[6] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[7] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[9] Dragomir R Radev, Eduard Hovy, and Kathleen McKeown. Introduction to the special issue on summarization. *Computational linguistics*, 28(4):399–408, 2002.

[10] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics.

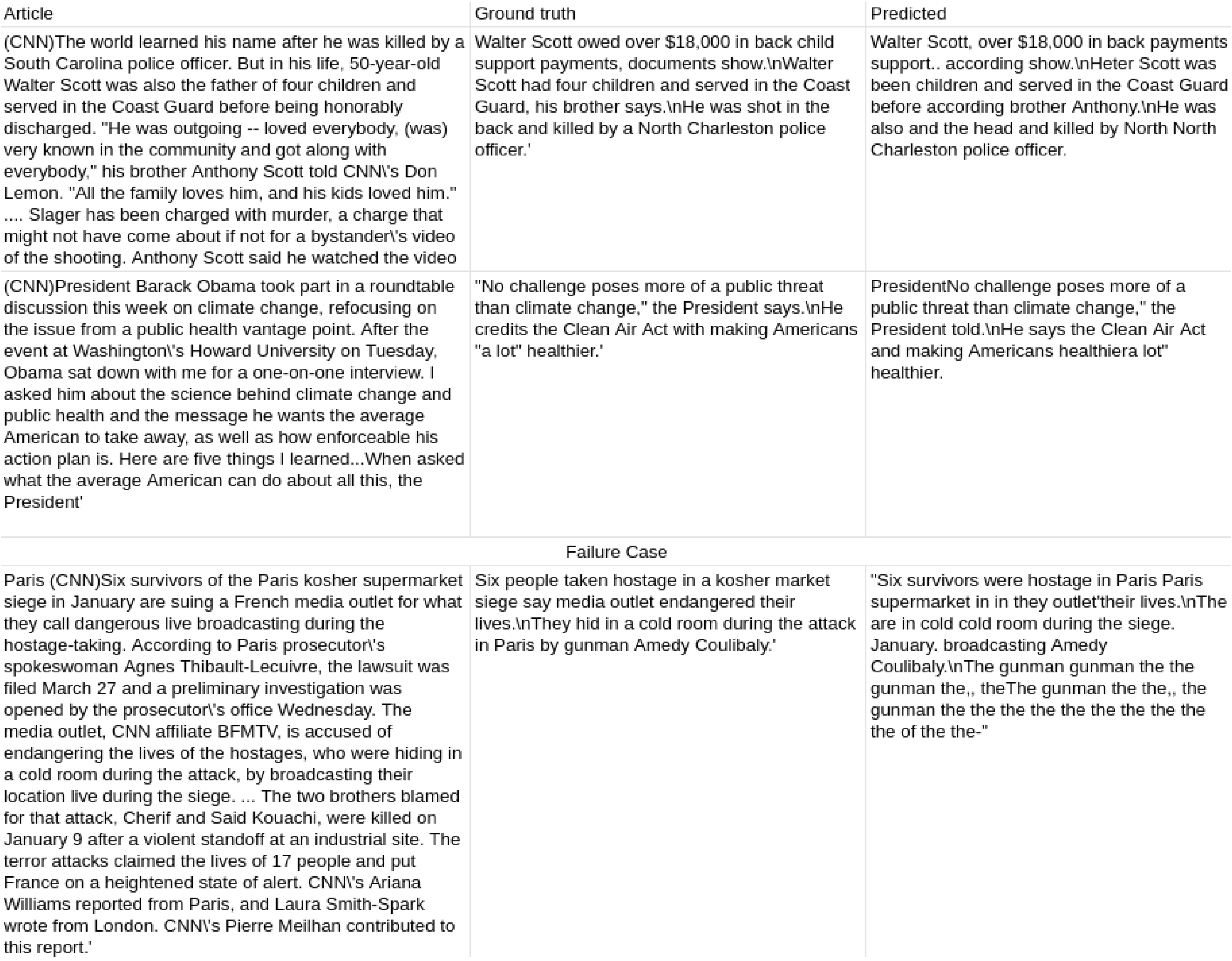
[11] Oguzhan Tas and Farzad Kiyani. A survey automatic text summarization. *PressAcademia Procedia*, 5(1):205–213, 2007.

[12] Amrisha Vaish, Tobias Grossmann, and Amanda Woodward. Not all emotions are created equal: the negativity bias in social-emotional development. *Psychological bulletin*, 134(3):383, 2008.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

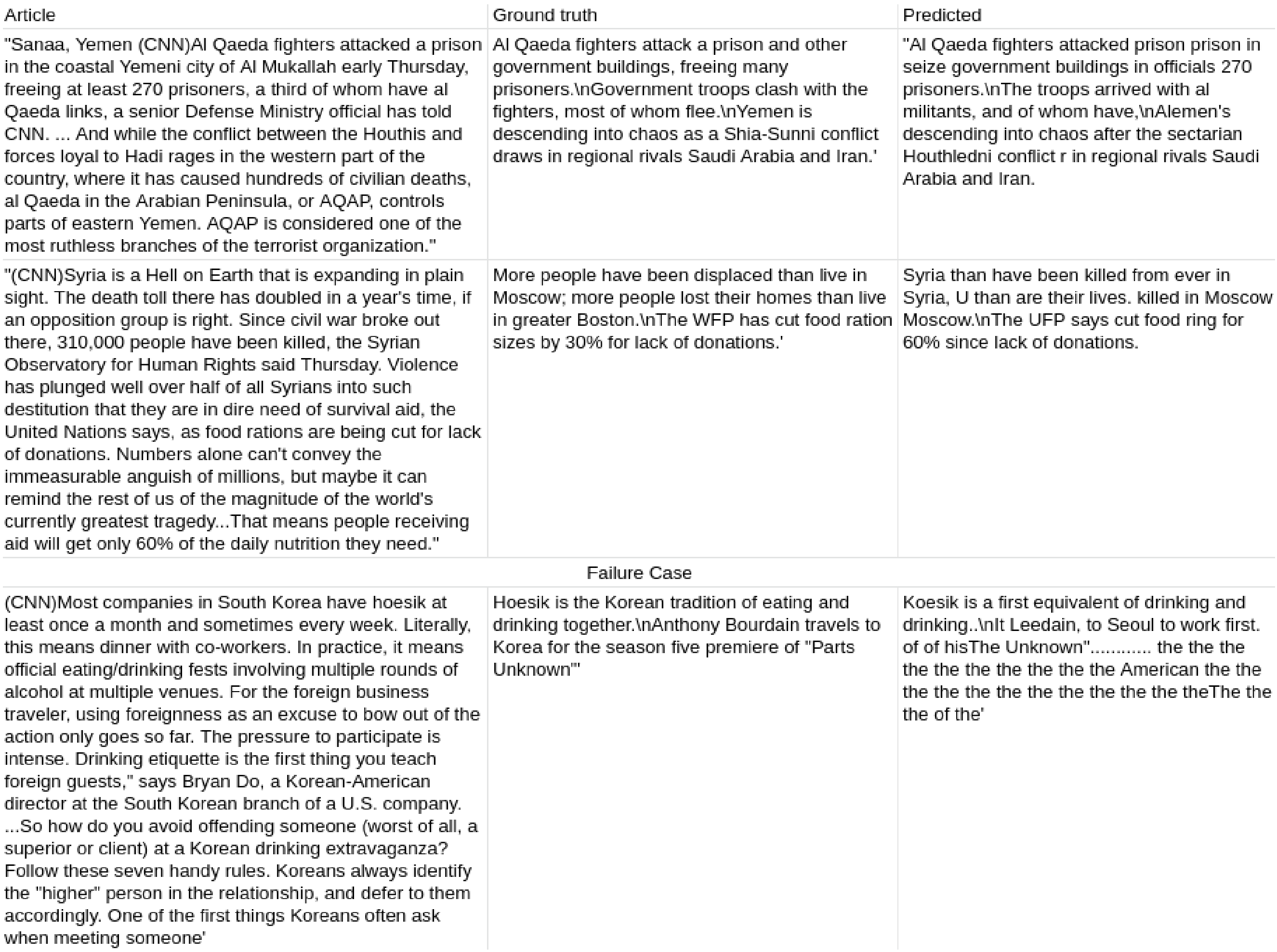
[14] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.

[15] Kelly W Zhang and Samuel R Bowman. Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis. *arXiv preprint arXiv:1809.10040*, 2018.



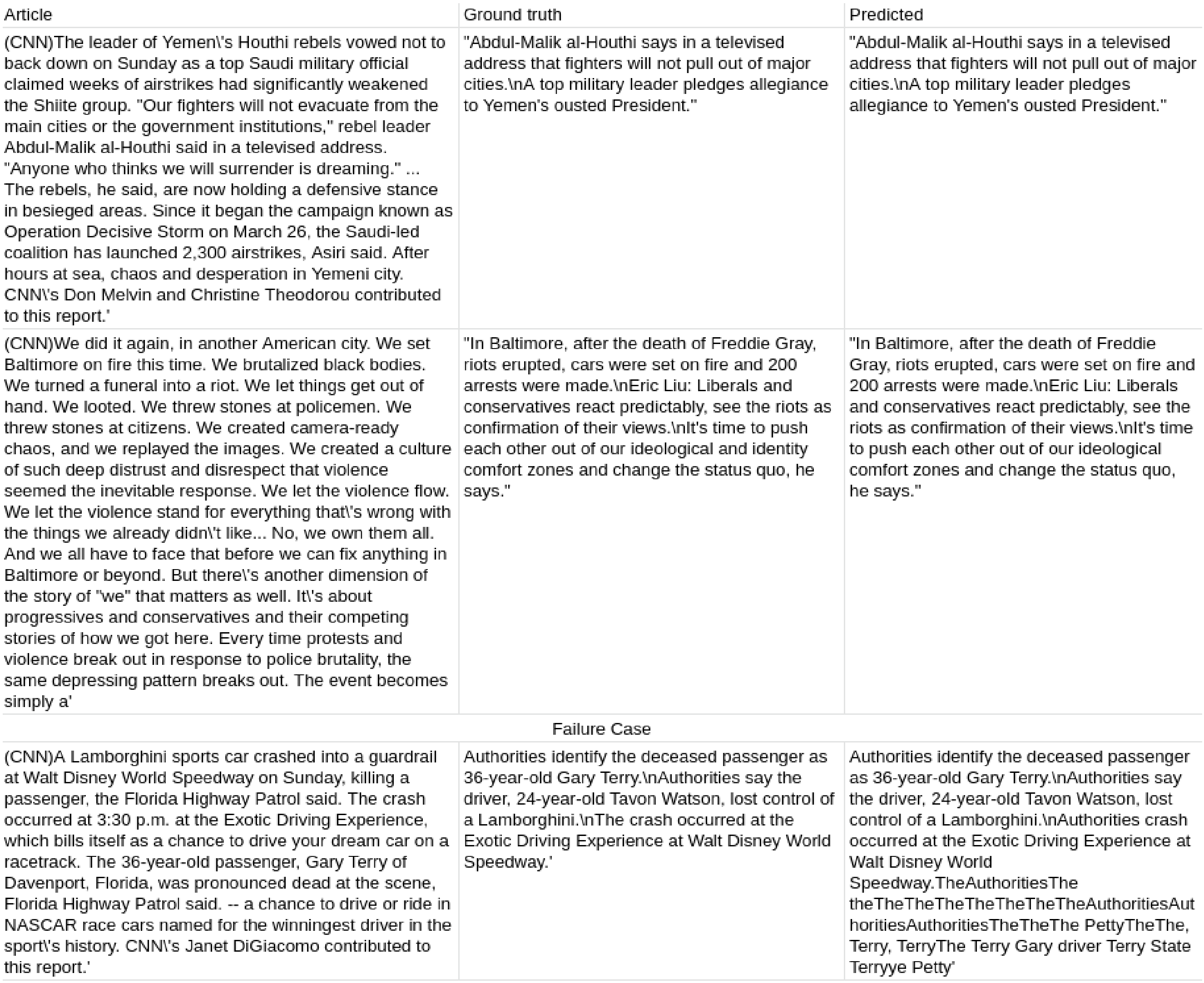
Article	Ground truth	Predicted
(CNN)The world learned his name after he was killed by a South Carolina police officer. But in his life, 50-year-old Walter Scott was also the father of four children and served in the Coast Guard before being honorably discharged. "He was outgoing -- loved everybody, (was) very known in the community and got along with everybody," his brother Anthony Scott told CNN\'s Don Lemon. "All the family loves him, and his kids loved him." .... Slager has been charged with murder, a charge that might not have come about if not for a bystander\'s video of the shooting. Anthony Scott said he watched the video	Walter Scott owed over $18,000 in back child support payments, documents show.\nWalter Scott had four children and served in the Coast Guard, his brother says.\nHe was shot in the back and killed by a North Charleston police officer.'	Walter Scott, over $18,000 in back payments support.. according show.\nHeter Scott was been children and served in the Coast Guard before according brother Anthony.\nHe was also and the head and killed by North North Charleston police officer.
(CNN)President Barack Obama took part in a roundtable discussion this week on climate change, refocusing on the issue from a public health vantage point. After the event at Washington\'s Howard University on Tuesday, Obama sat down with me for a one-on-one interview. I asked him about the science behind climate change and public health and the message he wants the average American to take away, as well as how enforceable his action plan is. Here are five things I learned...When asked what the average American can do about all this, the President'	"No challenge poses more of a public threat than climate change," the President says.\nHe credits the Clean Air Act with making Americans "a lot" healthier.'	PresidentNo challenge poses more of a public threat than climate change," the President told.\nHe says the Clean Air Act and making Americans healthiera lot" healthier.
	Failure Case	
Paris (CNN)Six survivors of the Paris kosher supermarket siege in January are suing a French media outlet for what they call dangerous live broadcasting during the hostage-taking. According to Paris prosecutor\'s spokeswoman Agnes Thibault-Lecuivre, the lawsuit was filed March 27 and a preliminary investigation was opened by the prosecutor\'s office Wednesday. The media outlet, CNN affiliate BFMTV, is accused of endangering the lives of the hostages, who were hiding in a cold room during the attack, by broadcasting their location live during the siege. ... The two brothers blamed for that attack, Cherif and Said Kouachi, were killed on January 9 after a violent standoff at an industrial site. The terror attacks claimed the lives of 17 people and put France on a heightened state of alert. CNN\'s Ariana Williams reported from Paris, and Laura Smith-Spark wrote from London. CNN\'s Pierre Meilhan contributed to this report.'	Six people taken hostage in a kosher market siege say media outlet endangered their lives.\nThey hid in a cold room during the attack in Paris by gunman Amedy Coulibaly.'	"Six survivors were hostage in Paris Paris supermarket in in they outlet'their lives.\nThe are in cold cold room during the siege. January. broadcasting Amedy Coulibaly.\nThe gunman gunman the the gunman the,, theThe gunman the the,, the gunman the the the the the the the the the the of the the-"

Figure 10. MSE equal weights generated captions (Article has been truncated)



Article	Ground truth	Predicted
"Sanaa, Yemen (CNN)Al Qaeda fighters attacked a prison in the coastal Yemeni city of Al Mukallah early Thursday, freeing at least 270 prisoners, a third of whom have al Qaeda links, a senior Defense Ministry official has told CNN. ... And while the conflict between the Houthis and forces loyal to Hadi rages in the western part of the country, where it has caused hundreds of civilian deaths, al Qaeda in the Arabian Peninsula, or AQAP, controls parts of eastern Yemen. AQAP is considered one of the most ruthless branches of the terrorist organization."	Al Qaeda fighters attack a prison and other government buildings, freeing many prisoners.\nGovernment troops clash with the fighters, most of whom flee.\nYemen is descending into chaos as a Shia-Sunni conflict draws in regional rivals Saudi Arabia and Iran.'	"Al Qaeda fighters attacked prison prison in seize government buildings in officials 270 prisoners.\nThe troops arrived with al militants, and of whom have,\nAlemen's descending into chaos after the sectarian Houthledni conflict r in regional rivals Saudi Arabia and Iran.
"(CNN)Syria is a Hell on Earth that is expanding in plain sight. The death toll there has doubled in a year's time, if an opposition group is right. Since civil war broke out there, 310,000 people have been killed, the Syrian Observatory for Human Rights said Thursday. Violence has plunged well over half of all Syrians into such destitution that they are in dire need of survival aid, the United Nations says, as food rations are being cut for lack of donations. Numbers alone can't convey the immeasurable anguish of millions, but maybe it can remind the rest of us of the magnitude of the world's currently greatest tragedy...That means people receiving aid will get only 60% of the daily nutrition they need."	More people have been displaced than live in Moscow; more people lost their homes than live in greater Boston.\nThe WFP has cut food ration sizes by 30% for lack of donations.'	Syria than have been killed from ever in Syria, U than are their lives. killed in Moscow Moscow.\nThe UFP says cut food ring for 60% since lack of donations.
Failure Case		
(CNN)Most companies in South Korea have hoesik at least once a month and sometimes every week. Literally, this means dinner with co-workers. In practice, it means official eating/drinking fests involving multiple rounds of alcohol at multiple venues. For the foreign business traveler, using foreignness as an excuse to bow out of the action only goes so far. The pressure to participate is intense. Drinking etiquette is the first thing you teach foreign guests," says Bryan Do, a Korean-American director at the South Korean branch of a U.S. company. ...So how do you avoid offending someone (worst of all, a superior or client) at a Korean drinking extravaganza? Follow these seven handy rules. Koreans always identify the "higher" person in the relationship, and defer to them accordingly. One of the first things Koreans often ask when meeting someone'	Hoesik is the Korean tradition of eating and drinking together.\nAnthony Bourdain travels to Korea for the season five premiere of "Parts Unknown"'	Koesik is a first equivalent of drinking and drinking..\nIt Leedain, to Seoul to work first. of of hisThe Unknown"............ the the the the the the the the the the the American the the the the the the the the the the the theThe the the of the'

Figure 11. Weighted MSE generated captions (Article has been truncated)



Article	Ground truth	Predicted
(CNN)The leader of Yemen\'s Houthi rebels vowed not to back down on Sunday as a top Saudi military official claimed weeks of airstrikes had significantly weakened the Shiite group. "Our fighters will not evacuate from the main cities or the government institutions," rebel leader Abdul-Malik al-Houthi said in a televised address. "Anyone who thinks we will surrender is dreaming." ... The rebels, he said, are now holding a defensive stance in besieged areas. Since it began the campaign known as Operation Decisive Storm on March 26, the Saudi-led coalition has launched 2,300 airstrikes, Asiri said. After hours at sea, chaos and desperation in Yemeni city. CNN\'s Don Melvin and Christine Theodorou contributed to this report.'	"Abdul-Malik al-Houthi says in a televised address that fighters will not pull out of major cities.\nA top military leader pledges allegiance to Yemen's ousted President."	"Abdul-Malik al-Houthi says in a televised address that fighters will not pull out of major cities.\nA top military leader pledges allegiance to Yemen's ousted President."
(CNN)We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. We created a culture of such deep distrust and disrespect that violence seemed the inevitable response. We let the violence flow. We let the violence stand for everything that\'s wrong with the things we already didn\'t like... No, we own them all. And we all have to face that before we can fix anything in Baltimore or beyond. But there\'s another dimension of the story of "we" that matters as well. It\'s about progressives and conservatives and their competing stories of how we got here. Every time protests and violence break out in response to police brutality, the same depressing pattern breaks out. The event becomes simply a'	"In Baltimore, after the death of Freddie Gray, riots erupted, cars were set on fire and 200 arrests were made.\nEric Liu: Liberals and conservatives react predictably, see the riots as confirmation of their views.\nIt's time to push each other out of our ideological and identity comfort zones and change the status quo, he says."	"In Baltimore, after the death of Freddie Gray, riots erupted, cars were set on fire and 200 arrests were made.\nEric Liu: Liberals and conservatives react predictably, see the riots as confirmation of their views.\nIt's time to push each other out of our ideological comfort zones and change the status quo, he says."
Failure Case		
(CNN)A Lamborghini sports car crashed into a guardrail at Walt Disney World Speedway on Sunday, killing a passenger, the Florida Highway Patrol said. The crash occurred at 3:30 p.m. at the Exotic Driving Experience, which bills itself as a chance to drive your dream car on a racetrack. The 36-year-old passenger, Gary Terry of Davenport, Florida, was pronounced dead at the scene, Florida Highway Patrol said. -- a chance to drive or ride in NASCAR race cars named for the winningest driver in the sport\'s history. CNN\'s Janet DiGiacomo contributed to this report.'	Authorities identify the deceased passenger as 36-year-old Gary Terry.\nAuthorities say the driver, 24-year-old Tavon Watson, lost control of a Lamborghini.\nThe crash occurred at the Exotic Driving Experience at Walt Disney World Speedway.'	Authorities identify the deceased passenger as 36-year-old Gary Terry.\nAuthorities say the driver, 24-year-old Tavon Watson, lost control of a Lamborghini.\nAuthorities crash occurred at the Exotic Driving Experience at Walt Disney World Speedway.TheAuthoritiesThe theTheTheTheTheTheTheTheAuthoritiesAuthoritiesAuthoritiesTheTheThe PettyTheThe, Terry, TerryThe Terry Gary driver Terry State Terryye Petty'

Figure 12. Weighted KL generated captions (Article has been truncated)