Enhancing Summarization with Text Classification via Topic Consistency

Jingzhou Liu (✉) and Yiming Yang

Carnegie Mellon University
{liujingzhou,yiming}@cs.cmu.edu

Abstract. The recent success of abstractive summarization is partly due to the availability of large-volume and high-quality human-produced summaries for training, which are extremely expensive to obtain. In this paper, we aim to improve state-of-the-art summarization models by utilizing less expensive text classification data. Specifically, we use an eXtreme Multi-label Text Classification (XMTC) classifier to predict relevant category labels for each input document, and impose topic consistency in the system-produced summary or in the document encoder shared by both the classifier and the summarization model. In other words, we use the classifier to distill the training of the summarization model with respect to topical consistency between the input document and the system-generated summary. Technically, we propose two novel formulations for this objective, namely a multi-task approach, and a policy gradient approach. Our experiments show that both approaches significantly improve a state-of-the-art BART summarization model on the CNNDM and XSum datasets. In addition, we propose a new evaluation metric, CON, that measures the topic consistency between the input document and the summary. We show that CON has high correlation with human judgements and is a good complementary metric to the commonly used ROUGE scores.

Keywords: text summarization · extreme multi-label text classification · multi-task learning · policy gradient.

1 Introduction

Text summarization is the task of condensing a given document into a shorter piece of textual summary, which preserves the main contents of the input. Existing approaches can be characterized into two categories, namely extractive and abstractive. Extractive methods compose each summary by extracting a subset of sentences from the input document, and abstractive methods produce each summary based on an underlying generative model, where the output may include the words or phrases beyond the input text. Generally speaking, extractive summaries are more fluent and accurate, while abstractive summaries can be globally more coherent and versatile. This paper focuses on improving abstractive summarization.

Significant improvements have been made recently in abstractive summarization [29, 8, 30, 25, 4, 7, 10], thanks to the rapid development of neural sequence-to-sequence learning techniques [31], such as attention [1], copy mechanism [11, 12], coverage [32] and reinforcement training [28]. However, many of these successful methods heavily rely on the availability of large-volume and high-quality of human-annotated summaries for training, which are extremely expensive to obtain. Later proposed models try to address this issue by first pre-training on massive unannotated corpora and then fine-tuning on supervised data [27, 15, 39]. A less explored direction is to leverage less expensive supervised data for other natural language processing tasks, such as those for classification, tokenization, textual entailment, etc. In this paper, we focus on the effective use of large-volume labeled documents in the area of eXtreme Multi-label Text Classification [17] (XMTC) for improving the state-of-art models in abstractive summarization.

The XMTC task is to find each document the most relevant subset of labels from an extremely large space of categories. For example, a news article reporting on the 2020 US election is likely to be labeled "US politics", "US election", "Donald Trump", "Joe Biden", etc., by an XMTC classifier trained with labeled Wikipedia articles. Those predicted category labels can be naturally viewed as a topical summary of the input article at a relatively high granularity, without including all the detailed information. In other words, a good summary of the input document should preserve the semantic information presented by those category labels at the desirable granularity. Then, the related question for research is: if we have sufficient labeled data to train a high-quality XMTC classifier, how can we use such a classifier to improve the training of a generative model for abstractive summarization?

We answer the above question by enforcing topic consistency between the classifier-predicted category labels and the system-produced summary for the input document. The reasons of focusing on XMTC are its multi-label predictions per document and the huge label spaces in general, ensuring adequate topic coverage and diversity for representing documents and their summaries. Specifically, in this paper, we utilize a convolutional neural net-based, efficient XMTC model, namely XML-CNN [17] trained on a Wikipedia dataset.

We propose two novel approaches to imposing topic consistency on summarization models: (1) a multi-tasking approach, and (2) a policy gradient approach. In the multi-tasking approach (as shown in Figure 1), the pre-trained XMTC classifier is applied to the input document to predict its topic distribution on one hand, and an additional MLP added to the encoder of the summarization model also predicts a topic distribution based on the latent embedding of the document for summarization. The encoder and the MLP are jointly trained to produce similar distributions, in addition to the seq2seq generative summarization objective. Thus, the summarization model is guided by the XMTC classifier to capture the most essential information with respect to subject topics from the input document. In the policy gradient approach (as shown in Figure 2), we apply the same pre-trained XMTC classifier to both the input document and

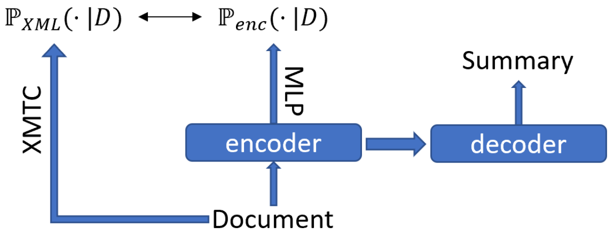

Fig. 1. Overview of the multi-tasking method. The XMTC classifier is applied to the input document to obtain document topic distribution, and the encoder of the summarization model is extended with an MLP as an induced classifier. The encoder induced classifier is trained to predict the document distribution, in addition to the normal sequence generation objective.

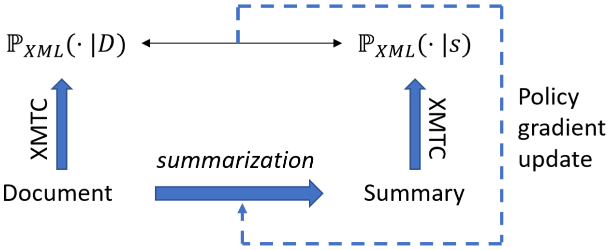

Fig. 2. Overview of the policy gradient method. The XMTC classifier is applied to both the input document and the system-produced summary, and then the predicted topic distributions are directly compared, sending training signal back to the summarization model via policy gradient.

the system-produced summary, and directly force these two distributions to be close to each other by minimizing a consistency reward and propagating back to the summarization model via policy gradient (Section 3). The two proposed approaches are generally applicable to any seq2seq-based summarization models. In this paper, we use a state-of-the-art summarization model, BART [15], as our base model, and we show that both of our proposed approaches produce significant improvements over the basic BART model (Section 4).

Additionally, for thorough evaluation, we compute topic consistency between an input document and a summary (topical information overlapping), and use this score as a complementary quality measure to the most commonly used n-gram overlapping-based ROUGE scores [16]. We show that CON scores have higher correlation with human evaluation scores than ROUGE, and therefore, should be informative when used as additional evaluation metric for methods comparison.

To summarize, the contributions of this paper are threefold:

1. We propose two novel approaches to imposing topic consistency to improve abstractive summarization models, by utilizing extreme multi-label text classification (XMTC).
2. Through experiments on two benchmark datasets, we show that our proposed methods significantly improve state-of-the-art summarization models.
3. We propose *consistency score*, CON, as a complementary evaluation metric to the commonly used metric, ROUGE score, and show that CON has higher correlation with human judgements than ROUGE.

2 Related Work

2.1 Abstractive Summarization

[29] and [8] were among the first to apply neural networks and attention mechanisms to abstractive summarization and since then substantial effort has been spent in this direction. [30] further incorporated pointer-generator and coverage mechanism [11] which allows directly copying words from the input document and keeping track of words that have been covered. [4] used multiple agents to represent the input documents with hierarchical attentions and trained the entire model end-to-end with reinforcement learning. [25] used reinforcement learning to directly optimize for ROUGE scores while keeping the produced summaries fluent and readable by mixing in cross-entropy objectives. [7] proposed a hybrid method that first extracts sentences using reinforcement learning, and then abstractively summarizes each extracted sentence into a summary.

As large-scale pre-training models such as BERT [9], roBERTa [19], and XLNet [37] were introduced, we have seen further improvements in abstractive summarization. [18] directly used pre-trained BERT encoders in summarization models. T5 [27], BART [15] and PEGASUS [39] pre-trained large-scale sequence-to-sequence models on massive unannotated corpora with specially designed tasks, and then further fine-tuned their models on downstream tasks (e.g. summarization) to achieve superior performances. [26] pre-trained a sequence-to-sequence model with future n-gram prediction to avoid overfitting on strong local correlations.

2.2 Combining Summarization and Text Classification

Summarization and text classification are two important and extensively studied tasks in natural language processing, and there are existing works trying to combine these two tasks. [3] proposed to utilize a multi-class (as opposed to multi-label) classifier to extract document representations, and to further use the predicted category to rank sentences in the document to compose extractive summaries. Other works combined sentiment classification with review summarization [36, 20, 5]: [36] proposed to jointly train a sentiment classifier and review summarizer by using semantic/sentiment/aspect attentions to better capture the review sentiment; [20] proposed an end-to-end model where a sentiment classifier is on top of the summarization model and jointly trained; [5] proposed a dual-view model where different sentiment classifiers are applied to reviews and summaries, for measuring sentiment inconsistency. All of these works focused on review summarization and preserving sentiment in the input reviews using (multi-class) sentiment classification, and they worked on datasets where both sentiment classification and review summarization supervision is available. To the best of our knowledge, this work is the first to utilize *external multi-label* text classifiers to capture *important topical information* to enhance abstractive summarization models.

2.3 Summarization Evaluation Metrics

Most commonly used metrics for evaluating text generation are based on counting n-gram overlapping between documents and summaries, such as ROUGE [16], BLEU [23] and METEOR [2]. However, it is known that these n-gram-based metrics have several deficiencies [40, 34, 14]. For example, n-gram-based metrics often fail to robustly match paraphrases due to string matching or heuristic matching, and n-gram-based metrics are usually insensitive to semantic errors and factual inconsistencies. To address these issues, several works have been proposed to improve or replace existing n-gram-based metrics. BERTScore [40] proposed to use contextual embeddings from BERT to compute similarity scores. [14] formulated the procedure of fact checking between input documents and summaries as a natural language inference problem. [34] further proposed to evaluate factual consistency in a question-answering framework: questions are generated based on the ground-truth summary, and then the QA system needs to answer the generated questions based on the system-produced summary. Our proposed consistency scores CON (Section 3.3), on the other hand, measures the topical information overlapping between a document and a summary, providing a cheap and efficient complementary metric to the most commonly used ROUGE scores.

3 Methodology

3.1 XMTC for Topic Distribution Prediction

Recall that extreme multi-label text classification (XMTC) is the problem of mapping each document to the subset of relevant categories in an extremely large space of predefined categories. Given document set D and category set L, we train the classifier $f(d) \to \{0,1\}^{|L|}$ for each $d \in \mathcal{D}$, where the size of category set L is usually extremely large, up to millions. As discussed in Section 1, the system-predicted categories for each document should capture important topical information about the input document, and provide useful guidance to the training of summarization models. To implement this idea, we utilize a representative convolutional neural net-based XMTC classifier, namely XML-CNN [17], which performs strongly on evaluation benchmarks and keeps a good balance among classification accuracy, model simplicity and computation scalability [38, 6].

An overview of the XML-CNN model is shown in Figure 3. It consists of a convolutional layer, a dynamic pooling layer, a bottleneck layer, and an output layer that predicts a probabilistic score for each label in the category space. Sorting and thresholding on the predicted scores allow us to select the top-ranking categories (e.g., 5∼20) for each input document and set the rest of the labels with zero scores. Then we normalize the topic scores to obtain a topic distribution conditioned on each input document. Let us denote the classifier by $C(\cdot)$, and then the document-conditioned topic scores for the document (d) and the system-produced summary (s) are $C(d) \in \mathcal{R}^{|L|}$ and $C(s) \in \mathcal{R}^{|L|}$, respectively. Our goal in this paper is to impose topic consistency between $C(d)$ and $C(s)$.

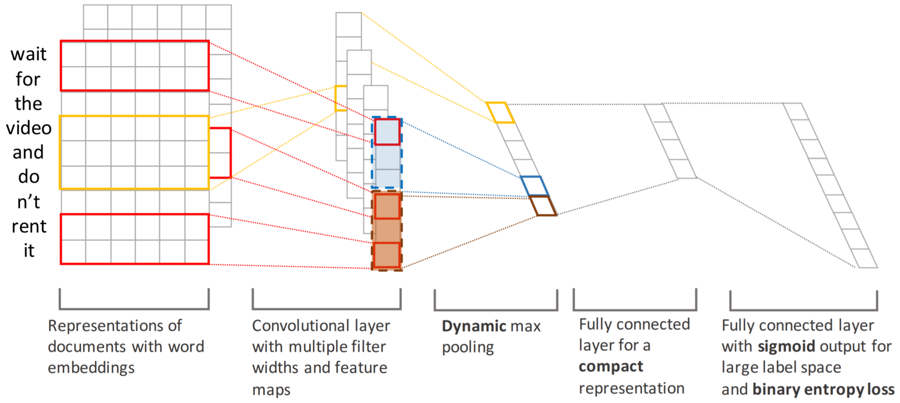
wait
for
the
video
and
do
n't
rent
it
Representations of documents with word embeddings
Convolutional layer with multiple filter widths and feature maps
Dynamic max pooling
Fully connected layer for a compact representation
Fully connected layer with sigmoid output for large label space and binary entropy loss

Fig. 3. Overview of XML-CNN, figure taken from [17].

Table 1. Data Statistics of Wiki-30K: L is the total number of class labels, $\bar{L}$ is the average number of label per document, $\tilde{L}$ is the average number of documents per label, $\bar{W}$ and $\hat{W}$ are the average number of words per document in the training/testing set.

#training	#testing	#vocab	L	$\bar{L}$	$\tilde{L}$	$\bar{W}$	$\hat{W}$
12,959	5,992	100,819	29,947	18.74	8.11	2,247	2,210

In the experiments of this paper, we train the XML-CNN classifier on a benchmark dataset named Wiki-30K [17], which is a collection of Wikipedia pages with human-curated category labels. Table 1 shows the dataset statistics. The label set consists of the 30K relatively popular category labels of Wikipedia articles. We choose this dataset to train our XML-CNN model because of its broad topic coverage and sufficient diversity, and hence rich enough for representing the topics in a wide range of documents and their summaries at various granularity levels. We also found the XML-CNN trained on this dataset can produce accurate predictions, which is also a desired property for the classifier to succeed in classification-enhanced summarization models.

3.2 Enhancing Summarization with Topic Consistency

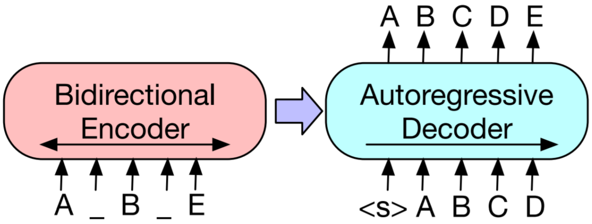

Fig. 4. Overview of BART, figure taken from [15].

Our base summarization model is BART [15], a state-of-the-art summarization model, as shown in Figure 4. It is a sequence-to-sequence model, where the encoder and decoder are transformers [33]. BART is pre-trained on large corpora using a variety of transformation tasks, including token masking/deletion/infilling, etc. To perform summarization task, BART is further fine-tuned on summarization datasets [15]. We start from a pre-trained BART model (without fine-tuning), and propose two approaches to imposing topic consistency on the system-produced summaries.

Multi-Tasking A natural idea is to formulate this as a multi-task learning problem, by enabling the summarization model also to be able to generate a topic distribution and making this generated topic distribution to be close to the one generated from the input document.

To achieve this, we add an additional MLP onto the encoder of the summarization model as an induced classifier. Specifically, the output of the encoder is a set of hidden states, which are averaged into a single vector, and then a two-layer MLP with separate sigmoid activation output is added on top of it. This two-layer MLP has the same shape as the last two layers of XML-CNN (Figure 3), and its parameters are initialized from our trained XML-CNN classifier (later both the encoder-decoder and the addition MLP are trainable). Now the encoder of the summarization model and the additional MLP together can be viewed as an induced classifier, denoted by $C_{enc}(\cdot)$.

Then, given training data of (document, summary) pairs $\{(d_i, s_i)\}_1^n$, we can define the consistency loss as

$$\mathcal{L}_{con} = \frac{1}{n}\sum_{i=1}^{n} ||C(d_i)^{(k)} - C_{enc}(d_i)||_2^2, \tag{1}$$

where $C(d_i)^{(k)}$ means only preserving the k largest entries in $C(d_i)$ and setting the rest to zeros. The reason we only focus on the top k labels is that since the label set L is huge (30K labels in Wiki30K), the predicted scores for most tail labels are quite noisy and unreliable, and that we only want the system-produced summary to capture the most salient and confident topics from the input document. The influence of different choices of k is investigated in Section 4.3.

Denote the regular cross-entropy loss from pairs of document and ground-truth summary by $\mathcal{L}_{xe}$, the final mixed loss is

$$\mathcal{L}_{MT} = \mathcal{L}_{xe} + \alpha \mathcal{L}_{con}. \tag{2}$$

Here α is a tunable hyper-parameter. In practice, the cross-entropy objective requires more training, and $\mathcal{L}_{con}$ starts to provide useful training signals after the encoder is moderately trained. Therefore, in our experiments, we first train the model only using $\mathcal{L}_{xe}$ for 3 epochs, and then further train the entire model using the mixed loss in Equation 2 until convergence.

Policy Gradient Another approach is to directly impose the constraints on the system-produced summaries using policy gradient method. As shown in Figure 2, a summary is generated from the summarization model without modification, and then the same, fixed XML-CNN classifier is applied to both the input document and the produced summary to obtained document topic distribution and summary topic distribution, denoted by $\mathbb{P}_{XML}(\cdot|d)$ and $\mathbb{P}_{XML}(\cdot|s)$, respectively. Based on these, a reward function measuring topic consistency between the input document and summary can be defined as

$$r_{con}(d, s) = -dist(\mathbb{P}_{XML}(\cdot|d), \mathbb{P}_{XML}(\cdot|s)), \tag{3}$$

where $dist(\cdot, \cdot)$ is a distance function. Topic distributions $\mathbb{P}_{XML}(\cdot|d)$ and $\mathbb{P}_{XML}(\cdot|s)$ come from the output of the XML-CNN model $C(d)$ and $C(s)$ (by normalizing the predicted scores), and we implement the reward function as

$$r_{con}(d, s) = -\frac{1}{k}||C(d)^{(k)} - C(s)^{(k)}||_2^2, \tag{4}$$

where the superscription (k) indicates preserving only the k largest entries and setting the rest to zeros.

The operation of feeding generated summaries to the XMTC classifier is non-differentiable, and so the training signals from the topic distribution mismatch cannot be back-propagated to the summarization model. To address this, we adopt the reinforcement learning-based training method. The summarization model can be viewed as a policy for generating summaries, denoted by p_θ, where θ is the parameters of the summarization model. Using the REINFORCE [35] algorithm, we minimize the RL loss function

$$\min_\theta \mathcal{L}_{rl} = -\mathbb{E}_{d \sim D} \mathbb{E}_{s \sim p_\theta(d)} [r_{con}(d, s)], \tag{5}$$

and its one sample derivative approximation is

$$\nabla_\theta \mathcal{L}_{rl} = -(r_{con}(d, s^*) - b_e) \nabla_\theta \log p_\theta(d, s^*). \tag{6}$$

Here s^* is the best summary generated by p_θ using beam search, and b_e is a baseline estimator to reduce variance during training. Following [24] and [28], baseline b_e is set to be $r_{con}(d, s^a)$, where s^a is obtained by top-1 greedy decoding.

Following [25, 24], we also optimize the cross-entropy generation loss in addition to the REINFORCE loss, to maintain the fluency and readability of the generated summaries. The cross-entropy loss is denoted by $\mathcal{L}_{xe}$, same as in Equation 2, and the final mixed loss is

$$\mathcal{L}_{PG} = \mathcal{L}_{xe} + \beta \mathcal{L}_{rl}. \tag{7}$$

Here β is a tunable hyper-parameter, and similar to the multi-tasking method, the summarization model is trained using only $\mathcal{L}_{xe}$ for 3 epochs first before using the mixed loss. The pseudo-code for policy gradient training is shown in Algorithm 1.

Algorithm 1: Pseudo-code for policy gradient approach training.

```
Data: Training samples $T = \{(d_i, s_i)\}_1^n$, XMTC classifier $C$
Result: Summarization model $p_\theta$
Train $p_\theta$ with cross-entropy objective for $t_1$ epochs on $T$;
for $t_2$ epochs do
    for training sample $(d, s) \in T$ do
        $g_{xe} \leftarrow$ gradient from cross-entropy objective;
        generate summary $s^* \leftarrow p_\theta(d)$ using beam search;
        generate baseline summary $s^a \leftarrow p_\theta(d)$ using greedy decoding;
        compute reward $r_{con}(d, s^*)$ and baseline $b_e = r_{con}(d, s^a)$ (Eq 4);
        $g_{rl} \leftarrow$ gradient from RL objective (Eq 6);
        update $\theta$ with $g_{xe} + \beta g_{rl}$;
return $p_\theta$
```

3.3 Topic Consistency as an Additional Evaluation Metric

The idea of measuring the consistency between the input document and summary itself can be utilized to evaluate summary quality. Specifically, we define *consistency score* (CON) for a document-summary pair as

$$CON(d, s) = \frac{1}{k} ||C(d)^{(k)} - C(s)^{(k)}||_2. \tag{8}$$

This formulation is similar to the consistency reward defined in Equation 4. Only the largest k entries in the document and summary topic distributions are preserved to reduce noise, and then l_2 distance is computed. Intuitively, lower consistency scores, i.e., more shared topics between the document and summary, indicate better summary quality.

Recently, the most commonly used n-gram overlapping-based summarization evaluation metrics (such as ROUGE score) have been criticized [25, 40, 14, 34] for not being able to robustly match paraphrases, under-penalizing small word or ordering changes which lead to serious hallucinations (more details in Section 2.3), and several alternative metrics have been proposed [34, 40, 14]. Compared to these newly proposed metrics (which are based on question-answering, BERT, etc.), our proposed consistency score, CON, processes coarser information, but its underlying XML-CNN model is more efficient and more robust. It measures topical information overlapping between a document and a summary, and serves as a cheap and efficient complementary metric to the traditional n-gram overlapping-based ROUGE scores.

4 Experiments

4.1 Datasets

We conduct experiments on two benchmark datasets (statistics in Table 2).

Table 2. Dataset Statistics.

Dataset	# Sample Split	# Words (doc)	# Sents (doc)	# Words (summary)	# Sents (summary)
CNNDM	287,227/13,368/11,490	651.9	30.0	52.6	3.6
XSum	204,045/11,332/11,334	431.1	19.8	23.3	1

CNN/Daily Mail [13] is a large-scale dataset consisting of news articles from the CNN and Daily Mail news publishers, and the reference summaries are human-generated highlights associated with these news articles. We do not anonymize named entities and follow the pre-processing procedure in [30]. It has been reported in previous literature that lead bias (important information of an article is covered in the first few sentences) is quite serious [30].

XSum [21] is an extreme news summarization dataset consisting of BBC articles and accompanying single-sentence summaries, where the single-sentence summaries are professionally written introductory sentences. Compared to the CNN/Daily Mail dataset, XSum is less biased toward extractive methods in that gold summaries in XSum contain significantly more novel n-grams than in CNN/Daily Mail, and so XSum is a more abstractive dataset.

4.2 Comparing Methods

- **Refresh** [22] is a strong supervised extractive method. It treats extractive summarization task as a sentence ranking task, and globally optimizes ROUGE score through a reinforcement learning objective.
- **Pointer-Generator** [30] is a supervised abstractive method that is based on the sequence-to-sequence framework, and is able to directly copy phrases from the input document via copy attention mechanism.

- **BertSumExtAb**s [18] is a supervised hybrid method that uses pre-trained large BERT models as document encoder, and then further fine-tunes the model on supervised summarization datasets.
- **Pegasus** [39] pre-trains a transformer-based sequence-to-sequence model on large-scale unannotated news corpora with masked language model and gap sentences generation, which are designed for the summarization task. The pre-trained model is then fine-tuned on supervised summarization datasets. Pegasus has two versions: Pegasus$_{base}$, which has 12-layer transformer encoder/decoder (24 layers in total), and Pegasus$_{large}$, which has 16-layer transformer encoder/decoder (32 layers in total).
- **BART** [15] is also a pre-trained sequence-to-sequence model fine-tuned on summarization dataset (more details in Section 3.2). It has 12-layer transformer encoder/decoder (24 layers in total).

4.3 Main Results

Table 3. Main Results on CNNDM and XSum Datasets. ROUGE scores and consistency scores (CON) are reported. 3 sentences are extracted for Refresh, instead of 4 [41]. For XMTC-enhanced models (row 6&7) and computing CON scores, k is set to 5. * indicates statistically significantly better than the best results in the same column through row 0-5.

		CNNDM				XSum			
ID	Method	R-1	R-2	R-L	CON	R-1	R-2	R-L	CON
0	Lead	40.5	17.7	36.7	0.25	16.3	1.6	12.0	0.48
1	Refresh	41.3	18.4	37.5	0.23	-	-	-	-
2	Pointer-Generator	39.5	17.3	36.4	0.27	29.7	9.2	23.2	0.36
3	BertSumExtAbs	42.1	19.6	39.2	0.23	38.8	16.5	31.3	0.34
4	Pegasus$_{base}$ (24 layers)	41.8	18.8	38.9	0.22	39.8	16.6	31.7	0.34
5	BART (24 layers)	44.2	21.3	40.9	0.20	45.1	22.3	37.3	0.31
	Our Methods								
6	BART-Enhanced-MT (24 layers)	44.8*	21.6	41.6*	0.18*	45.7*	22.8*	37.8*	0.29*
7	BART-Enhanced-PG (24 layers)	**45.2***	**22.0***	**42.1***	**0.14***	**45.9***	**23.0***	**38.1***	**0.25***

Main results are shown in Table 3. Statistical significance tests are conducted (including later experiments, Table 5&6) using paired t-test on summary-level with p value at the 5% level. Lead (row 0) is an extractive baseline that simply takes the first three sentences in CNNDM and the first sentence in XSum as summaries. Row 6&7 are our proposed methods that enhance a pre-trained BART summarization model with XMTC predictions via multi-tasking and policy gradient, respectively. Comparing row 3-7 with row 0-2, it is clear that models utilizing large pre-trained models achieve superior performances. Pegasus (row 4) and BART (row 5) are two state-of-the-art methods. They have the same architecture, and the differences are their pre-training corpora and pre-training tasks. Comparing row 6&7 with BART (row 5), both of our two proposed methods surpass their base BART model. Specifically, policy gradient method (row 7) leads to larger improvements than multi-tasking method (row 6), showing that directly imposing topic consistency via policy gradient is more effective than indirect multi-tasking method. Moreover, improvement of our methods (row 6&7) over BART (row 5) is slightly larger in CNNDM than in XSum, probably due to the fact that summaries in CNNDM are longer than in XSum (52.6 vs 23.3) and so summaries in CNNDM may carry more topical information.

The larger version of Pegasus (row 4) contains 36 layers, and here we did not run our model using 36 layers due to computation resource limitations. The results of Pegasus$_{large}$ (36 layers) on CNNDM and XSum in ROUGE scores are (44.2,21.5,41.1) and (47.2,24.6,39.3), respectively [39]. Our proposed 24-layer PG model (row 7) is better than Pegasus$_{large}$ on CNNDM but worse on XSum. However, our proposed methods of incorporating XMTC are orthogonal directions of improving summarization to better pre-training techniques (e.g. BART, Pegasus). Our proposed methods can be generally applied to enhance any sequence-to-sequence-based summarization model.

Table 4. Experiments on CNNDM with different k.

Method		CNNDM		
		R-1	R-2	R-L
Bart		44.2	21.3	40.9
Bart-Enhanced-PG	k=1	44.10	21.35	40.96
	k=5	45.22	21.96	42.10
	k=10	**45.26**	**22.04**	**42.13**
	k=20	45.13	21.68	41.85

In previous experiments, k (the number of preserved labels) is set to 5. To investigate the influence of k on the final performances, additional experiments on CNNDM using BART-Enhanced-PG are conducted with different k's. As shown in Table 4, when $k = 1$, the performance is much worse than larger k's, and is close to BART, indicating that extracting only one topic label from the document and summary does not provide much useful training signal and lead to any improvement. When $k = 5, 10, 20$, the performances are close, with $k = 20$ being slightly worse. Based on these observations, our proposed approaches to enhancing summarization models are quite robust when k is between 5 and 20 (i.e., adequate but not too noisy topic information is leveraged).

4.4 Human Evaluation

To further confirm the effectiveness of our proposed methods, we conduct human evaluations on summaries produced by BART, BART-Enhanced-MT and BART-Enhanced-PG. We sampled 200 documents from the test set, and AMT workers were asked to rate the system-produced and ground-truth summaries' quality given the input original document, on a 1-5 scale (higher is better). Each summary was shown to three different workers, whose ratings were averaged

Table 5. Human evaluation. Ratings are on a 1-5 scale, higher is better. * indicates statistically significantly better than ground-truth.

Summary	CNNDM	XSum
Ground-Truth	3.2	3.5
Bart	3.1	3.3
Bart-Enhanced-MT	3.2	3.3
Bart-Enhanced-PG	**3.4***	**3.4**

into a final rating. The results are shown in Table 5. We can see that the human evaluation results roughly agree with the automatic evaluation results in Table 3. Notably, in CNNDM, the summaries produced by Bart-Enhanced-PG are statistically significantly better than ground-truth summaries.

Table 6. Absolute Pearson correlation coefficients between automatic metrics and human ratings. * indicates statistically significantly better than the second best.

Metric	CNNDM	XSum
ROUGE-1	0.33	0.18
ROUGE-2	0.21	0.12
ROUGE-L	0.30	0.11
CON	**0.42***	**0.21**

Topic Consistency To investigate the consistency scores CON (Section 3.3), we further compute Pearson correlation coefficients (per summary) between various automatic evaluation metrics in Table 3 and human ratings in Table 5. The results are shown in Tabel 6. Clearly in this experiment, CON scores match human judgements better than ROUGE scores (with statistical significance on CNNDM). It should be noted that CON scores only consider shared topical information between input documents and system-produced summaries, but not the summaries' readability, fluency, grammaticality, etc., and so conceptually CON is not suitable to be used alone, but rather as a complementary metric to n-gram overlapping based metrics such as ROUGE scores for comparing methods (e.g., when ROUGE scores of two methods are close).

5 Conclusions

In this paper, we propose to utilize the cheaper extreme multi-label text classification (XMTC) data to enhance abstractive summarization models, where it is much more expensive to obtain supervised data. We propose two methods to impose topic consistency on the input documents and system-produced summaries using an XML-CNN classifier trained on a Wikipedia dataset, namely a multi-tasking method and a policy gradient method. Both methods manage to significantly improve a state-of-the-art BART summarization model. We also propose consistency score, CON, for evaluating summary quality, and show that

CON has higher correlation with human judgements than the most commonly used ROUGE scores. As for related future research directions, we would like to investigate the effective use of topic hierarchies and labeled documents to improve topic-conditioned summarization and multi-granularity summarization.

Acknowledgments

We thank all the reviewers for their helpful comments. This work is supported in part by National Science Foundation (NSF) under grant IIS-1546329.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
3. Cao, Z., Li, W., Li, S., Wei, F.: Improving multi-document summarization via text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
4. Celikyilmaz, A., Bosselut, A., He, X., Choi, Y.: Deep communicating agents for abstractive summarization. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1662–1675 (2018)
5. Chan, H.P., Chen, W., King, I.: A unified dual-view model for review summarization and sentiment classification with inconsistency loss. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1191–1200 (2020)
6. Chang, W.C., Yu, H.F., Zhong, K., Yang, Y., Dhillon, I.S.: Taming pretrained transformers for extreme multi-label text classification. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3163–3171 (2020)
7. Chen, Y.C., Bansal, M.: Fast abstractive summarization with reinforce-selected sentence rewriting. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 675–686 (2018)
8. Chopra, S., Auli, M., Rush, A.M.: Abstractive sentence summarization with attentive recurrent neural networks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 93–98 (2016)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. 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). pp. 4171–4186 (2019)
10. Gehrmann, S., Deng, Y., Rush, A.M.: Bottom-up abstractive summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4098–4109 (2018)

11. Gu, J., Lu, Z., Li, H., Li, V.O.: Incorporating copying mechanism in sequence-to-sequence learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1631–1640 (2016)
12. Gulcehre, C., Ahn, S., Nallapati, R., Zhou, B., Bengio, Y.: Pointing the unknown words. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 140–149 (2016)
13. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: Advances in neural information processing systems. pp. 1693–1701 (2015)
14. Kryscinski, W., McCann, B., Xiong, C., Socher, R.: Evaluating the factual consistency of abstractive text summarization. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 9332–9346 (2020)
15. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880 (2020)
16. Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. pp. 150–157 (2003)
17. Liu, J., Chang, W.C., Wu, Y., Yang, Y.: Deep learning for extreme multi-label text classification. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 115–124 (2017)
18. Liu, Y., Lapata, M.: Text summarization with pretrained encoders. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3730–3740 (2019)
19. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
20. Ma, S., Sun, X., Lin, J., Ren, X.: A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 4251–4257 (2018)
21. Narayan, S., Cohen, S.B., Lapata, M.: Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1797–1807 (2018)
22. Narayan, S., Cohen, S.B., Lapata, M.: Ranking sentences for extractive summarization with reinforcement learning. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1747–1759 (2018)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
24. Pasunuru, R., Bansal, M.: Multi-reward reinforced summarization with saliency and entailment. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 646–653 (2018)

25. Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. In: International Conference on Learning Representations (2018)
26. Qi, W., Yan, Y., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., Zhou, M.: Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. pp. 2401–2410 (2020)
27. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**, 1–67 (2020)
28. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7008–7024 (2017)
29. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 379–389 (2015)
30. See, A., Liu, P.J., Manning, C.D.: 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). pp. 1073–1083 (2017)
31. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
32. Tu, Z., Lu, Z., Liu, Y., Liu, X., Li, H.: Modeling coverage for neural machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 76–85 (2016)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
34. Wang, A., Cho, K., Lewis, M.: Asking and answering questions to evaluate the factual consistency of summaries. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5008–5020 (2020)
35. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning **8**(3-4), 229–256 (1992)
36. Yang, M., Qu, Q., Shen, Y., Liu, Q., Zhao, W., Zhu, J.: Aspect and sentiment aware abstractive review summarization. In: Proceedings of the 27th international conference on computational linguistics. pp. 1110–1120 (2018)
37. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems. pp. 5753–5763 (2019)
38. You, R., Zhang, Z., Wang, Z., Dai, S., Mamitsuka, H., Zhu, S.: Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. Advances in Neural Information Processing Systems **32**, 5820–5830 (2019)
39. Zhang, J., Zhao, Y., Saleh, M., Liu, P.: Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In: International Conference on Machine Learning. pp. 11328–11339. PMLR (2020)
40. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2019)
41. Zheng, H., Lapata, M.: Sentence centrality revisited for unsupervised summarization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6236–6247 (2019)