Noised Consistency Training for Text Summarization

Junnan Liu, Qianren Mao, Bang Liu, Hao Peng, Hongdong Zhu, Jianxin Li
Beihang University, University of Montreal
{lujn,maoqr,penghao,lijx}@act.buaa.edu.cn, liubang@iro.umontreal.ca

ABSTRACT
Neural abstractive summarization methods often require large quantities of labeled training data. However, labeling large amounts of summarization data is often prohibitive due to time, financial, and expertise constraints, which has limited the usefulness of summarization systems to practical applications. In this paper, we argue that this limitation can be overcome by a semi-supervised approach: consistency training which is to leverage large amounts of unlabeled data to improve the performance of supervised learning over a small corpus. The consistency regularization semi-supervised learning can regularize model predictions to be invariant to small noise applied to input articles. By adding noised unlabeled corpus to help regularize consistency training, this framework obtains comparative performance without using the full dataset. In particular, we have verified that leveraging large amounts of unlabeled data decently improves the performance of supervised learning over an insufficient labeled dataset.

KEYWORDS
Text Summarization, Semi-supervised Learning, Consistency Training, Language Model, Data-augmentation.

1 INTRODUCTION
Automatic text summarization [15] is a challenging task which generates a condensed version of an input text that captures the original’s core meaning. A fundamental requirement of automatic text summarization is that it typically requires a lot of labeled data to work well. However, acquiring handcrafted labels is a costly process, which motivates research on methods that can effectively utilize abundant unlabeled data to improve text summarization performance. We aim to provide an empirical answer to the following research question: what is the efficient way to leverage a large amount of unlabeled data to address the weakness of insufficient annotation for text summarization tasks?

Towards this goal, we first revisit an effective semi-supervised method, consistency training [16], to address the above issues. The consistency training leverages voluminous unlabeled data and employs data augmentation methods to generate diverse and realistic noisy source text, forcing the model to be consistent with these noises. The consistency regularization of semi-supervised learning has been extensively studied on the classification problems, such as text classification [5], image recognition [17], automatic speech recognition [9]. However, it is still unclear how semi-supervised learning works on text summarization tasks.

We believe that a good summarization model should be robust to any articles’ small changes. No prior research has ever tested this or a similar idea, so we try to perform consistency training in text summarization tasks. We first transform unlabeled articles into original samples and noisy samples to augment the supervised learning for text summarization with a sizeable unlabeled corpus. Then, consistency learning is used to ensure similar semantics of articles can be mapped to the same or similar output in prediction distributions. Thus, consistent regularization of semi-supervised learning can take effect on supervised learning performance over insufficient labeled data.

We evaluate text summarization on two full labeled datasets: CNN/DailyMail [4], and BBC XSum [8]. We split the full dataset into unlabeled and labeled data, e.g., a small-scale labeled dataset and a large-scale unlabeled dataset. We use two data augmentation methods: back-translation and word replacement with TF-IDF, to transform unlabeled data into noise-injected unlabeled data. The semi-supervised learning framework’s joint optimization includes two parts: supervised learning using labeled data, unsupervised consistency learning using unlabeled data, and noise-injected unlabeled data. We show that consistency-regularization semi-supervised learning could lead to a competitive result on a dataset with partial labeled and partial unlabeled data, compared with the performance on a full labeled dataset. In particular, we empirically observed performance gains for consistency training framework, compared with supervised baselines only using the labeled corpus. In summary, the main contributions of our method are:

- Discussing the feasibility of consistency-regularization semi-supervised learning in automatic text summarization tasks.
- Extensive evaluations demonstrate that consistency training with unsupervised corpus could greatly improve the performance of the text summarization model on a limited dataset.

1.1 Related Work
Abstractive text summarization is typically based on sequence-to-sequence (seq2seq) neural networks. The emergence of pre-training models for seq2seq learning [12] has extensively promoted the development of sequence generation tasks. Rothe et al. [13] develops a Transformer based seq2seq model that is compatible with publicly available pre-trained BERT [2], GPT-2 [11], and RoBERTa [7]
checkpoints. These models result in new salient performances on single document text summarization.

In recent work, consistency regularization methods for semi-supervised learning [1] have been shown to work well on many classification tasks [16]. The consistency training methods regularize model predictions to be invariant to noise applied to unlabeled examples. Tarvainen and Valpola [14] prove that a model trained with noisy labeled data learns to give consistent predictions around labeled data points. Additionally, high-quality data augmentation methods [16] can replace traditional noise injection methods to improve consistency learning performance. Their work can match and even outperform purely supervised learning that uses affluent labeled data.

Though pre-trained seq2seq learning and consistency training have recently achieved impressive gains in specific tasks respectively, it remains unclear to what extent combining pre-trained language models with consistency training can be beneficial to text summarization tasks.

2 FRAMEWORK

The consistent regularization of semi-supervised learning leverages unlabeled data and employs data augmentation methods to inject noisy data, and then enforces the summarization model to regularize semi-supervised learning by encouraging consistent predictions. We provide more details on our framework in what follows.

2.1 Data Augmentation.

We first transform unlabeled articles into original samples and noisy samples to improve the supervised learning for text summarization with a large unlabeled corpus. We use two advanced augmentation methods: back-translation and word replacement with TF-IDF.

Back-translation. Back-translation [3] refers to translating an existing document \(x\) from its original language A to another language B, and then translating it back to A to obtain an augmented document \(\tilde{x}\). As stated by Yu et al. [18], back-translation can produce different expressions while retaining the original sentence’s semantics, thereby achieving significant performance improvements in question answering tasks. In our work, we use back-translation to rewrite documents without changing the original intention\(^2\).

Word replacement with TF-IDF. This method tends to replace words with low TF-IDF scores while keeping high TF-IDF values [16]. Specifically, it would like to preserve keywords whose TF-IDF values are usually higher and replace non-informative words with other non-informative words.

2.2 Backbone Model: BERT2BERT

We focus on leveraging BERT [2] for summarization because BERT is often used as a benchmark model. We use BERT2BERT implemented by Rothe et al. [13] as the backbone model of the summarization framework. The BERT2BERT uses the BERT checkpoint to initialize the inputs to a context. Additionally, it uses the BERT checkpoint as the decoder to get better text generation from this context. We tokenize text using WordPiece \(^3\) to match the BERT pre-trained vocabulary.

2.3 Consistency Learning

A key aspect of consistent training to work well is to add noise to the input data. One assumption as to why noise is beneficial is that it enforces local smoothness for this task. This assumption supports our proposal for the following framework to improve the performance of text summarization models. In the text summarization task, the model generates a summary based on the overall understanding of a document so that a good model should also be invariant to documents with similar intentions. Hence, consistency learning can be used to ensure similar semantics of documents to be mapped to the same or similar output in prediction distributions.

Figure 1 gives an overview of our consistency training framework for text summarization. The inputs of the framework are labeled texts \(x\), unlabeled texts \(x'\), and noise injected unlabeled texts \(\tilde{x}\). We use \(y'\) to denote the gold summaries of labeled texts. Then we use \(f_{\theta}\) to represent the distributions of data generated by the model, where \(\theta\) refers to the model’s parameters. Our framework can be summarized as follows:

Firstly, we feed the labeled text \(x_i\) into the model to get the distribution \(f_{\theta}(y|x_i)\) and calculate the supervised loss:

\[
\frac{1}{n} \sum_{i=1}^{n} \ell(y'_i, f_{\theta}(y|x_i)) \tag{1}
\]

We then add noises to the texts to perform the consistency training in text summarization task. We generate a noised version \(\tilde{x}\) of unlabeled texts \(x'\) using data augmentation methods. Both unlabeled texts and augmented unlabeled texts are fed to the summarization model, and then we get the output distribution of original unlabeled data \(f_{\tilde{\theta}}(y'|\tilde{x})\) and an additional noised version of augmented unlabeled data \(f_{\tilde{\theta}}(y|\tilde{x})\).

We then calculate the semi-supervised loss between unlabeled texts and augmented unlabeled texts.

\[
\frac{1}{m} \sum_{i=1}^{m} \ell(f_{\theta}(y'|x'_i), f_{\tilde{\theta}}(y|\tilde{x}_i)), \tag{2}
\]

where \(\tilde{\theta}\) is just a copy of the current parameters \(\theta\) indicating that the back-propagation of the gradient is truncated. We use KL divergence loss to perform consistency training:

Finally, we combine supervised loss and semi-supervised loss, and train the model \(\theta\), by minimizing the combined loss:

\[
\frac{1}{n} \sum_{i=1}^{n} \ell(y'_i, f_{\theta}(y|x_i)) + \frac{1}{m} \sum_{i=1}^{m} \ell(f_{\theta}(y'|x'_i), f_{\tilde{\theta}}(y|\tilde{x}_i)). \tag{3}
\]

3 EXPERIMENTAL SETUP

3.1 Datasets

We conducted experiments on two datasets: CNN/DailyMail [4], and BBC XSum [8]. CNN/DailyMail includes news articles and

\(^2\)We use WMT'19 English-German translation models (in both directions) to perform back-translation on each article provided by Facebook-FAIR \(^7\). This model is the state-of-the-art language translation model.

\(^3\)https://github.com/google-research/bert/blob/master/tokenization.py
corresponding extractive highlights. We used the standard splits [4] for training, validation, and testing, which contains training samples with 287,227 pairs, 13,368 pairs for the development, and 11,490 pairs for the test. BBC XSum's summaries provide a high level of abstraction. The model requires document-level inference and paraphrasing to generate them. It includes 226,711 news articles and corresponding one-sentence summaries. We followed the splits [8] for training, validation, and testing with 20,404, 11,332, 11,334 article-summary pairs, respectively.

We obtained both labeled and unlabeled data from the full dataset. Specifically, we divided parts of the original dataset into labeled data. For the rest data, we deleted the labels and treated them as unlabeled data. For a concise comparison, we set the proportion of labeled data to three versions: 30% labeled data + 70% unlabeled data, 50% labeled data + 50% unlabeled data, and 70% labeled data + 30% unlabeled data.

3.2 Training details

During training on CNN/DailyMail dataset, the documents were truncated to 512 tokens, and the length of the summaries was limited to 128 tokens. For the BBC XSum dataset, the documents and summaries were truncated to 512 tokens and 64 tokens, respectively. We used a batch size of 16 for labeled data by default. Generally, semi-supervised learning performs a larger batch size on unlabeled data than labeled data to make full use of large quantities of unlabeled texts, refer to [16]. We implemented different batch sizes for unlabeled data and found that using a batch size of 32 leads to better performance.

All models were trained for 50,000 steps on 3 Tesla V100 GPUs. The learning rate started at 2e-5 and decayed every 1000 steps. We also performed a linear warmup method to increase the learning rate smoothly from 0 to 2e-5 during 2000 steps at the beginning of training. In the procedure of decoding, we used beam search (size 4) and set the 2.0 for the length penalty. The length of predicted summaries was limited to 162 for CNN/DailyMail and 62 for XSum. We also used trigram blocking to reduce repetition [10].

4 RESULTS ANALYSIS

Our experiments address the following research questions.

- **RQ1**: How does our proposed consistency training framework on single document text summarization perform?
- **RQ2**: Which of the two data aggregation methods has a greater impact on the overall performance?
- **RQ3**: How does the main hyperparameter of batch size affect the model's performance?

4.1 Main results of summarization (RQ1)

**On CNN/DailyMail dataset.** We evaluated the summarization predictions by ROUGE [6] in this paper. Table 1 shows the results on the CNN/DailyMail dataset.

The first block of Table 1 contains the baseline for different proportions of labeled data. The second block includes the consistency training models' results. As shown, consistency training models obtain improvements in ROUGE-1 points from 37.85 to 39.06, ROUGE-2 points from 15.58 to 17.16, and ROUGE-L points from 25.07 to 26.51 compared with the baseline on 30% labeled data. In the case where 50% of the data is labeled data, the model using consistency training achieves +0.64 point improvement in ROUGE-1, +0.67 point improvement in ROUGE-2, and +0.60 point improvement in ROUGE-L compared with baseline trained on 50% labeled data. We conclude that:

- The consistent training model trained on the dataset with partially labeled data (30%) and partial unlabeled data (70%) could achieve competitive results, compared to the model trained on the full dataset (100%).
- The model trained on the dataset with insufficient labeled data (30%) and large augmented unlabeled data (70%) performs performance gains with a large margin (+1.58 on R2, for instance) compared to the model only trained using 30% labeled data. The results indicate that consistent training with large unlabeled data could improve the supervised model's performance trained on a small labeled corpus.

**On BBC XSum dataset.** We also conducted experiments to verify if consistent training would be equally effective on the abstractive BBC XSum dataset. It has one-sentence summaries and is more abstractive than the CNN/DailyMail dataset.

As shown in Table 2, the first block shows the baseline of the different proportioned labeled data on the XSum dataset. The results on the XSum dataset using consistency training are summarized in the second block. Unexpectedly, the model with consistency training trained by 70% labeled data and 30% unlabeled data achieves similar results over the model trained on the full dataset. The results on XSum represent that our consistency-training method is also effective in generating extreme abstractive summaries.

| Table 1: ROUGE F1 results on the CNN/DailyMail dataset. |
| --- | --- | --- | --- |
| Labeled | ROUGE-1 | ROUGE-2 | ROUGE-L |
| 30% | 37.85 | 15.58 | 25.07 |
| 50% | 39.14 | 17.22 | 26.62 |
| 70% | 39.81 | 17.95 | 27.32 |
| 100% | 40.03 | 18.27 | 27.61 |

| Consistency training |
| --- | --- | --- | --- |
| Labeled + Unlabeled | ROUGE-1 | ROUGE-2 | ROUGE-L |
| 30% + 70% | 39.06 | 17.16 | 26.51 |
| 50% + 50% | 39.78 | 17.89 | 27.22 |
| 70% + 30% | 39.95 | 18.13 | 27.52 |
We performed experiments to discuss the impact of different data augmentation methods on the performance of consistency training. As shown in Table 3, the models using back-translation for data augmentation are superior to models using TF-IDF replacement. The former models outperform the latter ones by about +0.7 ROUGE-L points gains, +0.75 ROUGE-2 points gains, and +0.61 ROUGE-1 points gains on average.

These results show that back-translation leads to a better performance than TF-IDF replacement. We guess that back-translation could add more diversity to the text. On the other hand, back-translation can maintain the global semantics of sentences so as to maintain the input distribution.

### 4.2 Data augmentation methods study (RQ2)

We performed experiments to discuss the impact of different data augmentation methods on the performance of consistency training. As shown in Table 3, the models using back-translation for data augmentation are superior to models using TF-IDF replacement. The former models outperform the latter ones by about +0.7 ROUGE-L points gains, +0.75 ROUGE-2 points gains, and +0.61 ROUGE-1 points gains on average.

These results show that back-translation leads to a better performance than TF-IDF replacement. We guess that back-translation could add more diversity to the text. On the other hand, back-translation can maintain the global semantics of sentences so as to maintain the input distribution.

### 4.3 Unlabeled data batch size study (RQ3)

Generally, the batch size of unlabeled data in semi-supervised learning should be larger than the batch size of labeled data in order to make full use of unlabeled data. We performed experiments to investigate whether a larger batch size leads to better performance in text summarization tasks. Figure 2 shows Rouge-L results for different batch sizes of unlabeled data. We find that setting the batch size to 32 is better than setting it to 16, but setting it to 64 is slightly worse than setting it to 32. Sub-figure (b) shows the trend of loss curves under different batch sizes when validation. The batch size of 32 obtains optimal convergence speed.

### 5 CONCLUSIONS

Our work demonstrates that it is practical to use unlabeled data to improve the abstractive summarization model’s performance. We present a new perspective on effectively using consistency training to improve supervised text summarization over an insufficiently labeled dataset. By substituting simple noise injection operations with advanced data augmentation methods such as back-translation, our method brings substantial improvements across datasets with partial labeled and partially unlabeled data under the same consistency training framework. Our method obtains comparative performance without using the full dataset. Future works include migrating our consistency training framework to other natural language generation tasks, such as Q&A and dialogue generation.

## REFERENCES

[1] Philip Bachman, Ouais Alsharif, and Doina Precup. 2014. Learning with Pseudo Ensembles. In NIPS. 3365–3373.
[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT. Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics. 4171–4186.
[3] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In EMNLP. 489–500.
[4] Karl Moritz Hermann, Tomás Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In NIPS. 1693–1701.
[5] Alexander G. Ororbia II, C. Lee Giles, and David Reitter. 2015. Learning a Deep Hybrid Model for Semi-Supervised Text Classification. In EMNLP. 471–481.
[6] Chin Yew Lin. 2010. Automatic evaluation of summaries. In Workshop on Text Summarization Branches Out at ACL. 2004.
[7] Yinan Liu, Myle Ott, Nanfan Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692 (2019).
[8] Shashi Narayan, Shayan B. Cohen, and Mirella Lapata. 2018. Don’t Give Me the Details, Just the Summary? Topic-Aware Convolutional Neural Networks for Extreme Summarization. In ACL. 1797–1807.
[9] Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le. 2020. Improved Noisy Student Training for Automatic Speech Recognition. In INTERSPEECH ISCA. 2817–2821.
[10] Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A Deep Reinforced Model for Abstractive Summarization. In ICLR. OpenReview.net.
[11] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).
[12] Prajit Ramachandran, Peter J. Liu, and Quoc V. Le. 2017. Unsupervised Pretraining for Sequence to Sequence Learning. In EMNLP. 383–391.
[13] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging Pretrained Checkpoints for Sequence Generation Tasks. Trans. Assoc. Comput. Linguistics 8 (2020). 264–280.
[14] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In NIPS. 1195–1204.
[15] Juan-Manuel Torres-Moreno. 2014. Automatic Text Summarization. Wiley.
[16] Quize Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised Data Augmentation for Consistency Training. (2019).
[17] Quize Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. 2020. Self-Training With Noisy Student Improves ImageNet Classification. In CVPR.
[18] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. In ICLR.