Trainable Hard Negative Examples in Contrastive Learning for Unsupervised Abstractive Summarization

Haojie Zhuang[1], **Wei Emma Zhang**[1], **Chang George Dong**[1],
Jian Yang[2], **Quan Z. Sheng**[2]
[1]The University of Adelaide, Adelaide, Australia
[2]Macquarie University, Sydney, Australia
{haojie.zhuang,wei.e.zhang,chang.dong}@adelaide.edu.au
{jian.yang,michael.sheng}@mq.edu.au

Abstract

Contrastive learning has demonstrated promising results in unsupervised abstractive summarization. However, existing methods rely on manually crafted negative examples, demanding substantial human effort and domain knowledge. Moreover, these human-generated negative examples may be poor in quality and lack adaptability during model training. To address these issues, we propose a novel approach that learns *trainable* negative examples for contrastive learning in unsupervised abstractive summarization, which eliminates the need for manual negative example design. Our framework introduces an adversarial optimization process between a negative example network and a representation network (including the summarizer and encoders). The negative example network is trained to synthesize *hard* negative examples that are close to the positive examples, driving the representation network to improve the quality of the generated summaries. We evaluate our method on two benchmark datasets for unsupervised abstractive summarization and observe significant performance improvements compared to strong baseline models.

1 Introduction

Abstractive summarization is the task of generating concise summaries that potentially contain new phrases or sentences while preserving the core information of the source documents (See et al., 2017; Rush et al., 2015; Liu et al., 2022b; Nallapati et al., 2016). Abstractive summarization systems could be deployed in various applications such as news headline generation. Due to the challenge of collecting massive and high-quality parallel data (i.e., document-summary pairs) for training, it is increasingly important to study unsupervised abstractive summarization, which is especially valuable to uncommon domains and languages without sufficient labeled data (Liu et al., 2022a).

Document	... A new meme was born last night, once again at the expense of Miami Heat star forward LeBron James. The meme, #LeBroning, is flooding social media in response to James being carried off of the court in the waning minutes of the first game of the NBA Finals...Jordan famously played a game in the 1997 NBA Finals while suffering from influenza, winning the game ...
Negative Example	... A new meme was born last night, once again at the expense of Miami Heat star forward LeBron James. The meme, #LeBroning, is flooding social media in response to James being carried off of the court in the waning minutes of the first game of the NBA Finals...~~Jordan famously played a game in the 1997 NBA Finals while suffering from influenza, winning the game~~...
Gold Summary	Twitter and other social media exploded with mentions of #LeBroning following Thursday night's loss to the San Antonio Spurs. James claimed that he was experiencing cramping in last minutes of the game...

Table 1: An example (generated by deleting a random sentence from the source document) that is considered as a false negative example by all three annotators, since the deleted sentence is not important for the source document and summary.

Therefore, several models have been proposed for unsupervised summarization without the need for paired training data (Baziotis et al., 2019; Yang et al., 2020; Wang and Lee, 2018; Zhuang et al., 2022; Laban et al., 2020; Liu et al., 2022a; Schumann et al., 2020; Zhou and Rush, 2019). The recently proposed method SCR (Zhuang et al., 2022) applies contrastive learning in unsupervised abstractive summarization with outstanding performances. The model is trained to generate summaries and then to pull the summaries and positive examples in the semantic space while pushing away the summaries and negative examples, aiming to make the summaries preserve the key information. These negative examples in SCR are generated under some hand-crafted rules (e.g., in-

Findings of the Association for Computational Linguistics: EACL 2024, pages 1589–1600

sertion, deletion, replacement, entity swap). However, we notice that: (1) it requires human efforts and domain knowledge to design these rules. (2) the negative example generation rules in SCR could possibly generate low-quality negative examples or even false negative examples. For instance, as shown in Table 1, it may delete the non-essential or irrelevant sentences of the positive examples to create the negative examples, which would still be semantically the same as the positive examples. To further demonstrate this issue, we conduct a human evaluation to identify true or false negatives in SCR (details in Section 4.4) and show that only 25% are labeled as true negatives. These negative examples could confuse the model and hinder effective training by pushing apart the semantically similar examples. (3) Increasing the hardness of the negatives over the training process could improve the performance of contrastive learning (Wang et al., 2021). However, the rules in SCR are predefined and unchangeable, making the negative examples not adaptive to the model during the training. The adaptability would lead to a better and more robust match of positive pairs against negative pairs (Hu et al., 2021).

We are motivated to address these issues in (Zhuang et al., 2022) by taking advantage of *hard* negative examples, which are a type of *true negative* examples that are difficult to distinguish from the anchor (Robinson et al., 2021). Hard negative examples could help the model to capture the semantic similarity and thus improve the model performance (Xuan et al., 2020). Instead of using the hand-crafted rules, we propose to learn the ***trainable hard*** negative examples in an adversarial manner, where the negative examples are trained to be hard and diverse to improve the quality of the generated summaries. Specifically, we train two networks: (1) *Representation Network*, including the summarizer and encoders; and (2) *Negative Example Network* to synthesize hard negative examples for contrastive learning. Two networks are optimized alternatively. The representation network is optimized to minimize the contrastive loss, which minimizes the semantic distances between summaries and positive examples while maximizing that between summaries and negative examples. The negative example network is trained as "*counter-contrastive learning*" to maximize the contrastive loss by generating hard negative examples. The hard negative examples from the negative example network drive the representation network to improve the quality of summaries. Also, the synthesized hard negative examples could be adaptive to the representation network over the training.

The main contributions of this paper are summarized as follows,

- To the best of our knowledge, this work is the first attempt to study the problem of trainable hard negative examples in contrastive learning for unsupervised abstractive summarization.
- We propose a negative example network to generate hard negative examples adversarially in contrastive learning for unsupervised abstractive summarization.
- The experiment results demonstrate the effectiveness of our proposed methods, showing that the proposed method outperforms the current unsupervised summarization models in two benchmark datasets.

2 Related Work

2.1 Unsupervised Abstractive Summarization.

Recently, unsupervised approaches for abstractive summarization have been attracting increasing attention. Baziotis et al. (2019) and Wang and Lee (2018) learned to reconstruct the source inputs while the intermediate sequences serve as the output summaries. Two language models were proposed in Zhou and Rush (2019), where one enforced contextual matching and the other one targeted domain fluency. Schumann et al. (2020) used a hill-climbing algorithm for unsupervised sentence summarization with word extraction. Following Schumann et al. (2020), Liu et al. (2022a) trained an encoder-only non-autoregressive Transformer for summarization, which has also improved the inference efficiency. Yang et al. (2020) presented to pretrain with lead bias and fine-tuning on the target domain. Laban et al. (2020) aimed to optimize the summarization model for the important properties of a good summary: coverage, fluency and brevity. Three neural models were hence proposed to generate and evaluate the summaries. In Zhuang et al. (2022), a contrastive learning-based framework was proposed for unsupervised summarization, while the model was trained to output summaries that match the source documents semantically. We notice the negative examples generation strategies in Zhuang et al. (2022) are not always optimal and thus aim to improve the performance

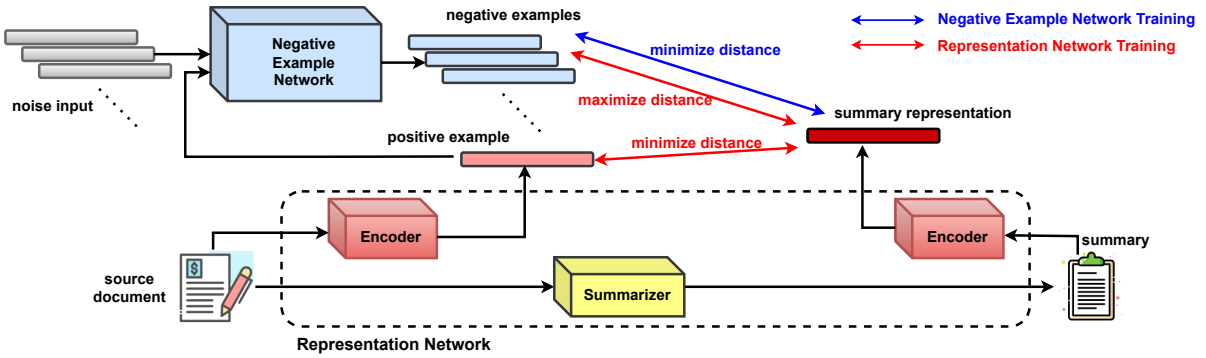

Figure 1: The overview of the proposed framework for trainable hard negative examples in contrastive learning for unsupervised abstractive summarization. The representation network includes the summarizer and the encoders. The negative example network is trained to generate hard negative examples for the representation network. The GAN loss (details in Section 3.3.3) has been omitted here for simplicity.

of contrastive learning for unsupervised abstractive summarization.

2.2 Hard Negative Examples.

Hard negative examples are shown to be effective in improving the performance of contrastive learning (Kalantidis et al., 2020; Xuan et al., 2020; Robinson et al., 2021). The authors in Kalantidis et al. (2020) uncovered that harder negative examples are helpful for better and faster learning, and thus proposed to synthesize hard negative examples in feature space for contrastive learning. For object detection, Lin et al. (2017) proposed a novel focal loss term to down-weight easy examples so that the model training would focus more on hard examples. Wang and Gupta (2015) used hard negative mining to learn more robust visual representations from unlabeled videos, where the top-K negative examples with the highest losses were selected for training. An Adversarial Contrast model was presented in Hu et al. (2021) to generate hard negative examples in an adversarial manner, which pushes the negative examples close to the positive queries. In Wang et al. (2021), the authors trained the model to generate hard negative examples for unpaired image-to-image translation with an adversarial loss. Inspired by Hu et al. (2021); Wang et al. (2021), we introduce the adversarial method to synthesize hard negative examples for contrastive learning in unsupervised abstractive summarization.

3 Methods

3.1 Preliminaries

We begin by having a brief introduction to the method SCR proposed in Zhuang et al. (2022), which applies contrastive learning for unsupervised abstractive summarization. In SCR, the summarizer first generates a summary given the source document, and then the model is trained with the contrastive encoder with contrastive loss:

$$l^{\hat{s}} = -\log \frac{\Big(\exp(\cos(\mathbf{v}_{\hat{s}}, \mathbf{v}_{c^+})/\tau)\Big)}{\Big(\exp(\cos(\mathbf{v}_{\hat{s}}, \mathbf{v}_{c^+})/\tau) + \sum_{c^-} \exp(\cos(\mathbf{v}_{\hat{s}}, \mathbf{v}_{c^-})/\tau)\Big)}, \quad (1)$$

where $\hat{s}$, c^+, c^- are the generated summary, positive example and negative example respectively; $\mathbf{v}_{\hat{s}}$, $\mathbf{v}_{c^+}$, $\mathbf{v}_{c^-}$ are their representation (encoded by the contrastive encoder) correspondingly; $\exp(\cdot)$ is the exponential function and $\cos(\cdot,\cdot)$ is the cosine similarity function; τ is the temperature.

The model is updated by minimizing the contrastive loss, which results in maximizing the similarity between the summaries and positive examples against the negative examples. The source document is considered as the positive example, while various human-designed strategies have been proposed to generate negative examples, such as sentence insertion, deletion, replacement, or entity swap of the source document. However, these strategies demand manual effort and can yield low-quality negative examples. Thus instead of using hand-crafted strategies, we aim to leverage the hard negative examples generated from a trainable network to perform more effective contrastive learning for unsupervised abstractive summarization.

3.2 The Proposed Model

As illustrated in Figure 1, the framework of the proposed model includes the representation network

$\mathcal{R}$ (including the summarizer and encoders) and negative example network $\mathcal{N}$. Two networks are optimized in an adversarial manner. Specifically, with a set of negative example representations from the negative example network $\mathcal{N}$, the representation network $\mathcal{R}$ is trained to minimize the semantic distance between the generated summaries and positive examples while maximizing that between the negative examples (as standard contrastive learning). Oppositely, the negative example network is optimized to maximize the contrastive loss while the representation network is fixed (as "counter-contrastive learning"). The adversarial training of two networks would drive the negative examples closer to the positive examples, which are more challenging and indistinguishable for the representation network. In the testing phase, we only use the summarizer to generate summaries given the source documents.

3.2.1 Representation Network

The representation network $\mathcal{R}$ consists of the summarizer and encoders. The summarizer aims to output the summary $\hat{s}$ given the source document d as input. The encoders generate the representations $\mathbf{v}_{\hat{s}}$, $\mathbf{v}_{c^+}$ for the summary $\hat{s}$ and positive example c^+ (also the source document d) respectively. Following Zhuang et al. (2022), we use the Transformer (Vaswani et al., 2017) with 6 layers and 8 attention heads (encoder and decoder) as the summarizer. For the encoders, we use a Transformer with 6 layers and 8 attention heads to encode the summary $\hat{s}$, while another Transformer with 12 layers and 12 attention heads to encode the source document d.

3.2.2 Negative Example Network

The negative example network $\mathcal{N}$ aims to generate hard negative examples that are close to the positive example, which is trained adversarially with the representation network $\mathcal{R}$. For each positive example, the negative example network $\mathcal{N}$ aims to output K negative example representations for contrastive learning. Concretely, the inputs for the negative example network are: (1) the positive example representation $\mathbf{v}_{c^+}$; (2) a random noise r_i $(1 \leq i \leq K)$ that are sampled from a normal distribution. The positive example representation input makes the negative examples instance-wise (highly related to the positive example), while the random noise input brings the randomness to have more diverse negative examples. We implement the negative example network as a three-layer MLP network to output as $\mathbf{v}^i_{c^-} = \mathcal{N}(\mathbf{v}_{c^+}; r_i)(1 \leq i \leq K)$.

3.3 Optimizaiton

The optimization objective for the model includes a contrastive loss for both representation network $\mathcal{R}$ (summaries and representations generation) and negative example network $\mathcal{N}$ (hard negatives generation); a diversity loss for $\mathcal{N}$ (diverse negatives generation); a GAN loss for $\mathcal{R}$ (summary quality improvement).

3.3.1 Contrastive Loss

The adversarial training of $\mathcal{R}$ and $\mathcal{N}$ could be formulated as a minimax optimization problem with Eq. (2) as follows,

$$\theta^*, \phi^* = \arg\min_{\theta} \max_{\phi} L^{con} \tag{2}$$

$$L^{con} = \mathbb{E}_d \left\{ -\log \frac{\left(\exp(\cos(\mathbf{v}_{\hat{s}}, \mathbf{v}_{c^+})/\tau)\right)}{\left(\exp(\cos(\mathbf{v}_{\hat{s}}, \mathbf{v}_{c^+})/\tau) + \sum_{i=1}^{K} \exp(\cos(\mathbf{v}_{\hat{s}}, \mathbf{v}^i_{c^-})/\tau)\right)} \right\} \tag{3}$$

where θ and ϕ are the parameters of $\mathcal{R}$ and $\mathcal{N}$. The $\mathbf{v}_{\hat{s}}$ and $\mathbf{v}_{c^+}$ in Eq. (3) are the function of θ, while $\mathbf{v}^i_{c^-}$ is the function of ϕ.

Due to the discrete output from the summarizer (part of the $\mathcal{R}$) that makes it difficult for gradient descent optimization, we use policy gradient (Sutton et al., 1999; Yu et al., 2017) as well as self-critical sequence training (Rennie et al., 2017) to update the summarizer. Hence, the loss for the summarizer could be re-written as:

$$L^{con}_G = -\mathbb{E}_{\hat{s}}[(-l^{\hat{s}} + l^{s_g}) \log p(\hat{s}_i | \hat{s}_1, \hat{s}_2, ...\hat{s}_{i-1})], \tag{4}$$

where $p(\hat{s}_i | \hat{s}_1, \hat{s}_2, ...\hat{s}_{i-1})$ is the output probability of the i-th token $\hat{s}_i$ conditioned on generated context $\{\hat{s}_1, \hat{s}_2, ...\hat{s}_{i-1}\}$. l^{s_g} is similar to $l^{\hat{s}}$ while replacing the $\hat{s}$ with s_g in Eq. 1, where s_g is the greedy-decoded output as a baseline (Wang and Lee, 2018; Zhuang et al., 2022).

3.3.2 Diversity Loss

Training the negative example network $\mathcal{N}$ with Eq. (3) could only lead to hard negative example generation. But these generated negative examples could possibly collapse to a single mode (Salimans et al., 2016; Wang et al., 2021). Therefore, we hope to synthesize diverse negative examples as well and

thus optimize the negative example network with another loss function by maximizing the difference of the negative example pairs, as follows,

$$L^{div} = -\|\mathbf{v}^i_{c^-} - \mathbf{v}^j_{c^-}\|, i \neq j \tag{5}$$

3.3.3 GAN Loss

Training only with the contrastive loss and diversity loss, the model is updated to generate a summary that could match the positive example semantically and keep away from the negative examples, while neglecting the writing quality (e.g., fluency, readability, etc) of the summary. To take it into account, we thus introduce another GAN loss (Goodfellow et al., 2014; Zhuang et al., 2022; Wang et al., 2021; Wang and Lee, 2018) for training (denoted as $\{L^{gan}_D, L^{gan}_G\}$), where the summarizer and a discriminator D are optimized adversarially. Specifically, the summarizer is trained to generate text that is similar to human-written text, while the discriminator tries to distinguish between text written by humans and summarizers. Following (Zhuang et al., 2022), we implement the discriminator as a Long short-term memory (LSTM) network (hidden size of 512), which is trained to output a score c_i at each time step t_i (denoted as $D(\cdot) = \{c_1, c_2, ...c_i, ...\}$). Also, to produce the human-written text s^r for the discriminator training, we extract the consecutive L sentences in each randomly sampled document from the dataset. We add the gradient penalty (Gulrajani et al., 2017) to the GAN loss for the discriminator, which could be formulated as follows,

$$\begin{aligned} L^{gan}_D = \mathbb{E}_{\hat{s}}[D(\hat{s})] - \mathbb{E}_{s^r}[D(s^r)] + \\ \lambda_D \mathbb{E}_{\bar{s}}[(||\nabla_{\bar{s}} D(\bar{s})||_2 - 1)^2], \end{aligned} \tag{6}$$

where $(||\nabla_{\bar{s}} D(\bar{s})||_2 - 1)^2$ is the gradient penalty (Gulrajani et al., 2017) (with the weight λ_D), and $\bar{s}$ is sampled from the linear interpolation between pairs of $\hat{s}$ and s^r.

Similarly, because of the non-differentiable problem of sampling, the GAN loss for the summarizer is re-written as:

$$L^{gan}_G = -\mathbb{E}_{\hat{s}}[(c_i - c_{i-1}) \log p(\hat{s}_i | \hat{s}_1, \hat{s}_2, ...\hat{s}_{i-1})], \tag{7}$$

where c_i and c_{i-1} are scores from the discriminator, and c_0 is set to 0 when $i = 1$.

Therefore, the overall loss for $\mathcal{R}$ and $\mathcal{N}$ is as follows (with the loss weights λ_{gan} and λ_{div}),

$$\begin{aligned} L_\theta &= L^{con} + \lambda_{gan} L^{gan}_G \\ L_\phi &= -L^{con} + \lambda_{div} L^{div} \end{aligned} \tag{8}$$

	CNN/DailyMail	Gigaword
length of document	781	29
length of summary	56	9
train/val/test	287k/13k/11k	3.8M/189k/2k

Table 2: The statistics of the datasets. The length is the average count of the token in documents or summaries.

4 Experiment

4.1 Experiment Settings

Datasets. To verify the effectiveness of our proposed methods, we conduct experiments on two widely used datasets: CNN/DailyMail (Nallapati et al., 2016; Hermann et al., 2015) and English Gigaword (Rush et al., 2015) datasets. We present the statistics of the datasets in Table 2. To have a fair comparison with other unsupervised abstractive summarization models, we only train our proposed model with the source documents, which means that our model has no access to any reference summary in the datasets.

Automatic Evaluation Metrics. We use the ROUGE F1 score (Lin, 2004) for evaluation, including uni-gram overlap (R1), bi-gram overlap (R2) and longest common subsequence (RL).

Baseline Models. We compare the 8 unsupervised summarization models with our proposed method: SEQ^3 (Baziotis et al., 2019); Adv-Reinforce (Wang and Lee, 2018); TED (Yang et al., 2020); Summary Loop (Laban et al., 2020); ContextualMatch (Zhou and Rush, 2019); HC_article_10 (Schumann et al., 2020); NAUS (Liu et al., 2022a); SCR (Zhuang et al., 2022). The model NAUS (Liu et al., 2022a) and HC_article_10 (Schumann et al., 2020) are proposed for unsupervised sentence summarization, hence they are only evaluated on the Gigaword dataset.

Training Details. We set the temperature τ and number of negative examples K in Eq. (3) as 1.0 and 128, respectively. The weight λ_D in Eq. (6) as 1.0, the weight λ_{gan} and λ_{div} in Eq. (8) as 0.85 and 1.0, respectively. The dimension of the $\mathbf{v}_{\hat{s}}$, $\mathbf{v}_{c^+}$ and $\mathbf{v}^i_{c^-}$ in Eq. 3 are 256. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1e-4. We also pretrain the proposed model (details in Appendix A). We run all experiments on a single Nvidia 3090 GPU.

4.2 Overall Results

The automatic evaluation results are shown in Table 3 (CNN/DailyMail) and Table 4 (Gigaword). Our method outperforms the strong baselines model on

Model	R1	R2	RL
SEQˆ3	23.24	7.10	22.15
Adv-Reinforce	35.51	9.38	20.98
TED	38.73	16.84	35.40
Contextual-Match	14.25	3.10	10.87
Summary Loop	37.70	14.80	34.70
SCR	39.06	17.43	37.12
Our method	**41.10**	**18.98**	**37.63**

Table 3: The experimental results on CNN/DailyMail (with 95% confidence interval). The bold scores represent the best performance.

Model	R1	R2	RL
SEQˆ3	25.39	8.21	22.68
Adv-Reinforce	28.11	9.97	25.41
TED	25.58	8.94	22.83
Contextual-Match	26.48	10.05	24.41
SCR	28.10	**11.63**	24.14
HC_article_10	24.44	8.01	22.21
NAUS	**28.55**	9.97	25.78
Our method	**28.55**	10.43	**26.11**

Table 4: The experimental results on Gigaword (with 95% confidence interval). The bold scores represent the best performance.

both datasets: (1) On CNN/DailyMail, our proposed method achieves better performance than other baselines in terms of R1, R2 and RL. Compared to the model SCR that applies human-design strategies to generate negatives (Zhuang et al., 2022), our model has 2.04, 1.55 and 0.51 improvement in R1, R2 and RL respectively, which could demonstrate the effectiveness of our negative examples network. (2) On Gigaword, our proposed method surpasses other models in R1 (same as NAUS (Liu et al., 2022a)) and RL, while R2 is the second best among all models. The competitive overall performance demonstrates the effectiveness of our proposed method.

4.3 Ablation Study

To further understand our proposed method, especially the impact of each component, we conduct the ablation test by removing: (1) contrastive learning loss for the negative example network, denoted as "**w/o** $\boldsymbol{L^{con}}$" (2) diversity loss for the negative example network, denoted as "**w/o** $\boldsymbol{L^{div}}$" (3) GAN loss for the summarizer, denoted as "**w/o** $\boldsymbol{L^{gan}}$". Table 5 provides the ablation study results. Not surprisingly, our proposed method achieves the best performance with all the components. Removing either component will lead to a significantly worse performance, which verifies the importance of these components for improving the overall quality of the output summaries.

Removing Component	R1	R2	RL
w/o L^{con}	19.23	6.45	15.09
w/o L^{div}	22.20	9.01	20.14
w/o L^{gan}	28.08	11.29	24.10

Table 5: The ablation study results on CNN/DailyMail

w/o $\boldsymbol{L^{gan}}$. We observe that the result of w/o L^{gan} is the best in Table 5. We believe the main reason is the role of GAN loss. Training without the GAN loss would sacrifice the writing quality of the generated summaries (such as grammar errors, or being unreadable), but the summaries could possibly preserve the key information from the source documents due to effective contrastive learning. Thus the summaries might contain more keywords or phrases (e.g., name entity) and have a higher ROUGE score since the ROUGE metric compares the word (or phrase) overlap between the summaries and references.

w/o $\boldsymbol{L^{con}}$. From the results, w/o L^{con} performs worse than w/o L^{div}, which we believe is reasonable because w/o L^{con} (only with diversity loss) could only generate diverse but low-quality negative examples. Such negative examples could be unrelated and not able to effectively push the summaries close to the documents, which would lead to poor-quality output summaries.

w/o $\boldsymbol{L^{div}}$. Training without the diversity loss also results in an inferior performance compared to the full model. We believe the main reason is: more diverse negative examples would be more challenging and thus could perform more effective contrastive learning (Xuan et al., 2020; Wang et al., 2021; Kalantidis et al., 2020).

Moreover, we show a generated summary example under the ablation settings in Appendix B.

4.4 Negative Examples Analysis

4.4.1 False Negatives Issue

To verify the false negatives issue in SCR, We first conduct a human evaluation to identify false negatives by randomly sampling 100 negative examples that are generated using the same rules as SCR. Then three annotators are asked to label each example as true or false negative example given the

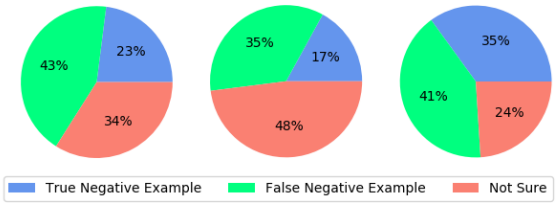

Figure 2: The statistical results of three annotators to identify true/false negative examples of SCR (Zhuang et al., 2022).

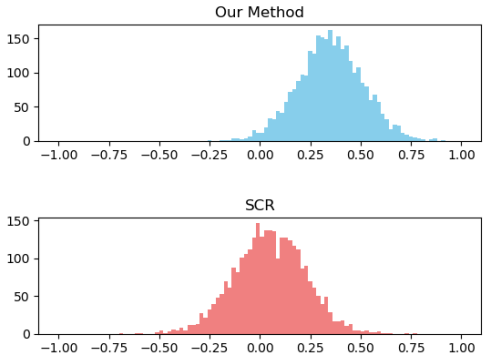

Figure 3: The similarity between summaries and negative examples in our method and SCR (Zhuang et al., 2022). The x-axis and y-axis are cosine similarity and frequency respectively.

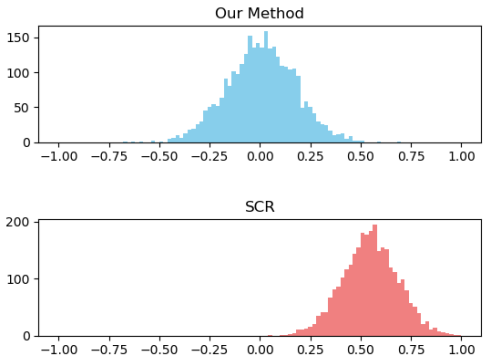

Figure 4: The diversity of the negative examples. Our method could generate more diverse negatives compared to SCR (Zhuang et al., 2022). The x-axis and y-axis are cosine similarity and frequency respectively.

source document and the reference summary. As shown in Figure 2, only 25% (on average) of negative examples generated by hand-crafted rules of SCR are considered as true negatives, and nearly 40% are labeled as false negatives. Furthermore, to explore whether our generated negative examples are more similar to the false negatives or true negatives. Specifically, we construct false negatives and true negatives in text space by: for each positive example (i.e., source document), we apply back-translation and synonym substitute to generate a semantically similar example as the false negative example. Moreover, we also obtain the true negative example by replacing the entities in the positive example (i.e., bringing factual errors). Then we use the representation network to encode these constructed false negatives and true negatives, which is followed by computing the cosine similarities of our generated negatives and the constructed false negatives (or true negatives). The experiment results show that 86.1% of our generated negative examples are more similar to the true negatives, indicating that our proposed method could effectively address the false negatives issue.

4.4.2 Similarity Between Summaries and Negative Examples

To understand the distribution of the trainable negative examples of our proposed method, we randomly sample 3,000 examples from the dataset and calculate the average cosine similarities between the summaries and negative examples generated by the negative example network. As the histogram shown in Figure 3, we could observe that the similarities in SCR (Zhuang et al., 2022) are mostly centered around 0, indicating that the negative examples are not pushed close enough to the summaries. The similarities in our method are much higher than in SCR, which we believe these negative examples generated by the negative example network are more challenging for the model to perform contrastive learning.

4.4.3 Diversity of the Negative Examples

Furthermore, to demonstrate the diversity of the negative examples, we also calculate the cosine similarities between the negative example pairs in SCR and our method. From the result in Figure 4, we find that the negative examples of our method are more diverse than SCR (as the negatives are less similar to each other). Our model could benefit from training with more diverse negative examples (details in Section 4.3).

Model	R1	R2	RL
Target Domain: Gigaword			
SCR	23.10	7.08	19.24
Our Method	**25.07**	**8.14**	**19.63**
Target Domain: CNN/DailyMail			
SCR	24.65	8.77	22.29
Our Method	**26.20**	**11.29**	**23.98**

Table 6: The experimental results of zero-shot summarization.

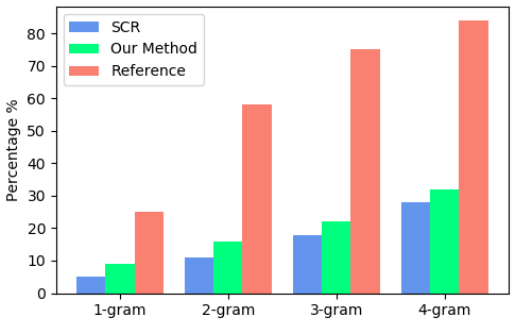

Figure 5: The statistical analysis of abstractiveness (novel N-grams in the summaries of different systems).

System	Rel	Coh	Con	Flu
SCR	2.81	3.08	2.90	3.13
Our method	3.44	3.57	3.47	**3.49**
Gold summary	**3.51**	**3.64**	**3.54**	3.41

Table 7: The human evaluation results on Rel (Relevance), Coh (Coherence), Con (Consistency) and Flu (Fluency).

4.5 Zero-shot Summarization

Following Zhuang et al. (2022), we conduct experiments to verify how well the model could be adapted to another dataset (or domain) by training the model on one dataset and then performing zero-shot summarization on another dataset. Specifically, we use the CNN/DailyMail dataset as the source domain to train our proposed model, followed by evaluating on the target domain Gigaword, and vice versa. As shown in Table 6, our proposed method outperforms SCR on both datsets, which demonstrates the advantages of the trainable negative examples over the hand-crafted rules in SCR for zero-shot summarization.

4.6 Abstractiveness

As we train our summarization model under the abstractive settings, we would like to understand how well our abstractive summarization model could avoid simply copying from the document. To analyze the model's abstractiveness, we count the novel words or phrases that are not present in the source documents. Specifically, we statistically analyze the novel N-gram ($N \in \{1, 2, 3, 4\}$) in the summaries (from SCR, our method and the reference summaries) on the CNN/DailyMail dataset and present the result in Figure 5. The statistical result indicates that our method could generate more abstractive summaries over the SCR model.

4.7 Human Evaluation

In addition to the automatic evaluation metrics, we also assess the quality of our model-generated summaries with human judgement. We randomly sample 100 examples from the CNN/DailyMail test set and then three expert annotators are invited to conduct the manual evaluation on the summary quality. They are presented with the source documents and the summaries from three systems (SCR (Zhuang et al., 2022), our method and gold summary). Following Fabbri et al. (2021); Kryscinski et al. (2019), each summary is evaluated across four dimensions: (1) **Relevance**: how good is the summary selecting the most important contents from the documents; (2) **Coherence**: the collective quality of all sentences in each summary; (3) **Consistency**: the factual consistency between the summary and the source document (hallucination content detection); (4) **Fluency**: the writing quality of individual sentences in the summary, such as being grammatically correct and readable for humans. Each summary was rated by three distinct judges and the final score is obtained by averaging the individual scores. The annotators rate each summary on a scale of 1 to 5 (with 1 being the worst and 5 being the best), while the final result of each system is the averaged score of the individual summary ratings. The average kappa score in our human evaluation is 0.84, which is able to indicate a strong inter-rater agreement.

We list the results in Table 7 and show that our method outperforms SCR (Zhuang et al., 2022) with higher human evaluation scores. Unsurprisingly, the gold summaries are ranked the best in relevance, coherence and consistency. Our proposed method is slightly better than the reference summaries in fluency. We showcase two examples in Appendix C to demonstrate the summary quality of our method.

5 Discussion and Conclusion

In the era of LLMs (Large Language Models), LLMs could generate high-quality summaries that are significantly preferred by humans (Pu et al., 2023; Zhang et al., 2023b). Why do we still study unsupervised summarization? We believe that LLMs (e.g. ChatGPT) are not suitable for all scenarios (e.g., confidential/sensitive data or domain, minority languages) (Huang et al., 2022; Patil et al., 2023; Kim et al., 2023; Zhang et al., 2023a), and thus it is still important to conduct research on training models for summarization tasks. In this paper, we have provided an unsupervised training strategy for summarization. Researchers or engineers could utilize our method to train the models on their own data (e.g. company confidential data, personal private data), domains (e.g. medical texts, legal documents), languages (e.g. minority language), where LLMs could not be used or might not be good enough. Since our method is unsupervised, there is no need for human-written summaries as references, thus significantly reducing human labor and costs in training. Besides, researchers could fine-tune their own LLMs or pretrained models (e.g. pretrained language models) using our method for better summarization performance. Our method could also be applied in some semi-supervised scenarios where limited human-written references are available.

To conclude our work, we explore and study the problem of trainable hard negative examples in contrastive learning for unsupervised abstractive summarization, and propose to train a negative example network and a representation network in an adversarial manner. The negative example network is optimized to generate high-quality and diverse hard negative examples for the representation network to generate better summaries and representations. Extensive experiments and analysis on two benchmark datasets demonstrate the effectiveness of our proposed method, as well as the significant advantages over the strong baseline models.

Limitations

While the output summaries of our proposed method obtain a high score in human evaluation, we observe the problem of factual inconsistency in some of the generated summaries. Summarization models are likely to output hallucination content that could not be entailed by the source document (Kryscinski et al., 2020; Maynez et al., 2020; Cao and Wang, 2021). This issue would limit our model to being reliable and trustworthy. Since our proposed method could be naturally included with other learning objectives (e.g., a factuality loss term), future research could extend our work with a factual consistency loss, which could improve the faithfulness and factuality of the output summaries. Besides, it is difficult to check what the negative examples look like in text space since it is even a more non-trivial task to generate texts given the representations. One possible solution is multi-task learning: to have an additional task of generating texts from representations during the training.

Acknowledgments

This work is supported by the Australian Research Council Early Career Industry Fellowship (IE230100119). The authors sincerely thank all the anonymous reviewers for their valuable comments and feedback.

References

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. SEQˆ3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 673–681.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, pages 391–409.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 2014 Conference on Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved training of wasserstein gans. In *Proceedings of the 2017 International Conference on Neural Information Processing Systems*, page 5769–5779.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman,

and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 2015 International Conference on Neural Information Processing Systems*, page 1693–1701.

Q. Hu, X. Wang, W. Hu, and G. Qi. 2021. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1074–1083.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, page 21798–21809.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models. *arXiv preprint arXiv:2307.01881*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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)*, pages 540–551.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst. 2020. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5135–5150.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Puyuan Liu, Chenyang Huang, and Lili Mou. 2022a. Learning non-autoregressive models from search for unsupervised sentence summarization. In *Proceedings of the 2022 Annual Meeting of the Association for Computational Linguistics*, pages 7916–7929.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022b. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 2022 Annual Meeting of the Association for Computational Linguistics*, pages 2890–2903.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 2016 SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Vaidehi Patil, Peter Hase, and Mohit Bansal. 2023. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. *arXiv preprint arXiv:2309.17410*.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive Learning with Hard Negative Samples. In *Proceedings of the 2021 International Conference on Learning Representations*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 2234–2242.

Raphael Schumann, Lili Mou, Yao Lu, Olga Vechtomova, and Katja Markert. 2020. Discrete optimization for unsupervised sentence summarization with word-level extraction. In *Proceedings of the 2020 Annual Meeting of the Association for Computational Linguistics*, pages 5032–5042.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 2017 Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1073–1083.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, page 1057–1063.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 2017 Conference on Advances in Neural Information Processing Systems*, page 5998–6008.

Weilun Wang, Wengang Zhou, Jianmin Bao, Dong Chen, and Houqiang Li. 2021. Instance-wise Hard Negative Example Generation for Contrastive Learning in Unpaired Image-to-Image Translation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14000–14009.

Xiaolong Wang and Abhinav Gupta. 2015. Unsupervised Learning of Visual Representations Using Videos. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802.

Yaushian Wang and Hung-Yi Lee. 2018. Learning to encode text as human-readable summaries using generative adversarial networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4187–4195.

Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. 2020. Hard negative examples are hard, but useful. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV*, page 126–142.

Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. 2020. TED: A pretrained unsupervised summarization model with theme modeling and denoising. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1865–1874.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the 2017 AAAI Conference on Artificial Intelligence*, page 2852–2858.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2023a. Mcˆ 2: A multilingual corpus of minority languages in china. *arXiv preprint arXiv:2311.08348*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023b. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Jiawei Zhou and Alexander Rush. 2019. Simple unsupervised summarization by contextual matching. In *Proceedings of the 2019 Annual Meeting of the Association for Computational Linguistics*, pages 5101–5106.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2021. Leveraging lead bias for zero-shot abstractive news summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1462–1471.

Haojie Zhuang, Wei Emma Zhang, Jian Yang, Congbo Ma, Yutong Qu, and Quan Z. Sheng. 2022. Learning from the source document: Unsupervised abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4194–4205.

A Model Pretraining

Following (Zhuang et al., 2022; Wang and Lee, 2018), we pretrain the representation network and negative example network respectively before jointly training the whole model. First, we take advantage of lead-bias (Zhu et al., 2021; Yang et al., 2020; Wang and Lee, 2018; Zhuang et al., 2022) to pretrain the summarizer to predict the first few sentences (or tokens) given the rest of the documents. Specifically, we set the first 3 sentences as the output references in CNN/DailyMail and the first 8 tokens in Gigaword. The pretraining would allow the model to infer the key information given the background content in the rest of the document, as well as to be trained as a simple language model. For other parts of the representation network and negative example network, we then use Eq. (8) for pretraining with the pre-trained summarizer.

B Example Summary of Ablation Study

We list an example summary under the ablation settings (w/o L^{con}, w/o L^{div} or w/o L^{gan}) in Table 9. All summaries generated by our model (under ablation settings) miss some key information. As we expected, w/o L^{gan} might generate some unreadable phrases (e.g., "running the CNN") that make humans difficult to understand. We also find that w/o L^{con} generates some inconsistent content (e.g., "gives away a mistake"), which is not supported by the source document. This example demonstrates the importance of three components.

C Case Study

We showcase two example summaries in Table 8. As shown in Example 1, our method could cap-

Example 1	
Document	... Bring your own beaker, goblet or vase and slurp it up. 7-Eleven is hosting the first Bring-Your-Own-Cup Slurpee Day at United States stores from 11 a.m. to 7 p.m. Saturday to kick off-peak Slurpee season... can fill their "cup" of choice for $1.49, the average cost of a medium Slurpee... The promotion isn't to be confused with Free Slurpee Day, traditionally celebrated each July 11.
Gold Summary	Bring your own large "cup" for a $1.49 7-Eleven Slurpee. Any sanitary container less than 10 inches in diameter is fair game.
SCR	7-Eleven is hosting the first Bring-Your-Own-Cup Slurpee Day at U.S. stores. Customers can fill their cup of choice for $1.49 ... a 10-inch-diameter hole will rule out anything too large.
Our Method	7-Eleven is hosting the first Bring-Your-Own-Cup Slurpee Day, not to be confused with Free Slurpees Day each July 11. Customers can fill their own cup of choice for $1.49.
Example 2	
Document	... Lionel Messi's became Barcelona's highest goalscorer of all time after his hat-trick against Granada helped the Spanish champions close on La Liga leaders Real Madrid ... took his tally to 234 goals for the club to overtake Cesar Rodriguez, who had a prolific spell for Barca in the 1940s and '50s. Messi took his tally for the season to 54 in 45 games as Barca's 5-3 victory on Tuesday ... Elsewhere, Juventus booked a place in the Italian Cup final after Mirko Vucinic's extra-time strike sealed a 4-3 aggregate win over AC Milan...
Gold Summary	Lionel Messi becomes the highest scorer in Barcelona's history after hat-trick against Granada. Argentina striker's treble helps Barca to a 5-3 win to close on La Liga leaders Real Madrid. Messi's haul takes his tally to 234 for the club, past Cesar Rodriguez's long-held record. The 24-year-old has 54 goals for the Spanish and European champions this season.
SCR	Lionel Messi becomes Barcelona's highest goalscorer of all time with a hat-trick against Granada... Barca's 5-3 victory on Tuesday helps cut the gap to Real Madrid to five points.Juventus booked a place in Italian Cup final after beating AC Milan 4-3 on aggregate.
Our Method	Lionel Messi becomes Barcelona's highest goalscorer of all time with a hat-trick against Granada... takes his tally to 234 goals for the club to overtake Cesar Rodriguez. Messi takes his season tally to 54 goals in 45 games as Barca beat Granada 5-3. Juventus advanced to the Italian Cup final after 4-3 aggregate win over AC Milan.

Table 8: Example summary of our proposed method. The words with the same colors share the same information between documents and summaries.

Document	...About 20 hours after the Boston Marathon...Venezuelan native Maickel Melamed, who is battling muscular dystrophy, completed the 26.2 miles just before 5 a.m. Tuesday. A group of energized fans rallied for the 39-year-old as he walked down... His perseverance was celebrated by crowds at the marathon finish line Tuesday morning, and also by fans online...
Gold Summary	Maickel Melamed, who has muscular dystrophy, took part in the 2015 Boston Marathon. He completed the race 20 hours after the start. Despite rainy weather, fans and friends cheered for the 39-year-old.
w/o L^{gan}	... Maickel Melamed is battling muscular dystrophy. He completed the 26.2-mile race ... running the CNN ...
w/o L^{div}	... Maickel Melamed, who is battling muscular dystrophy ...
w/o L^{con}	... Maickel Melamed completed the in this year's marathon ... gives away a mistake ...

Table 9: An example summary of our model under different ablation settings. Words in green are content in poor quality.

ture the key information from the source document, such as "the event of 7-Eleven", while discarding the unimportant details, e.g., the container requirements of the event. In Example 2, our method also retains the most important content, e.g., "Messi becomes Barcelona's highest goalscorer overtaking Cesar Rodriguez, his season tally, Juventus' victory" from the source document, while even the gold summary misses "Juventus' victory". We also observe the newly generated phrases: in Example 1, our model outputs the phrase "their own", which is not found in the original document; in Example 2, our summarizer rewrites "Juventus booked a place in the Italian Cup final" as "Juventus advanced to the Italian Cup final". Last but not least, the example summaries show that our method could generate fluent and coherent text.