Improving Truthfulness of Headline Generation

Kazuki Matsumaru **Sho Takase** **Naoaki Okazaki**
{kazuki.matsumaru, sho.takase, naoaki.okazaki}@nlp.c.titech.ac.jp
Tokyo Institute of Technology

Abstract

Most studies on abstractive summarization report ROUGE scores between system and reference summaries. However, we have a concern about the *truthfulness* of generated summaries: whether all facts of a generated summary are mentioned in the source text. This paper explores improving the truthfulness in headline generation on two popular datasets. Analyzing headlines generated by the state-of-the-art encoder-decoder model, we show that the model sometimes generates untruthful headlines. We conjecture that one of the reasons lies in untruthful supervision data used for training the model. In order to quantify the truthfulness of article-headline pairs, we consider the textual entailment of whether an article entails its headline. After confirming quite a few untruthful instances in the datasets, this study hypothesizes that removing untruthful instances from the supervision data may remedy the problem of the untruthful behaviors of the model. Building a binary classifier that predicts an entailment relation between an article and its headline, we filter out untruthful instances from the supervision data. Experimental results demonstrate that the headline generation model trained on filtered supervision data shows no clear difference in ROUGE scores but remarkable improvements in automatic and manual evaluations of the generated headlines.

1 Introduction

Automatic text summarization aims at condensing a text into a shorter version while maintaining the essential information (Mani, 2001). Methods on summarization are broadly categorized into two approaches: *extractive* and *abstractive*. The former extracts important words, phrases, or sentences from a source text to compile a summary (Goldstein et al., 2000; Erkan and Radev, 2004; Mihalcea, 2004; Lin and Bilmes, 2011). In contrast, the latter involves more complex linguistic operations (e.g., abstraction, paraphrasing, and compression) to generate a new text (Knight and Marcu, 2000; Clarke and Lapata, 2008). Until 2014, abstractive summarization had been less popular than extractive one because of the difficulty of generating a natural text. However, research on abstractive summarization has attracted a lot of attentions recently with the advances on encoder-decoder models (Rush et al., 2015; Takase et al., 2016; Zhou et al., 2017; Cao et al., 2018a; Song et al., 2019; Wang et al., 2019).

English Gigaword (Graff and Cieri, 2003; Napoles et al., 2012) is a representative dataset for abstractive summarization. Rush et al. (2015) regarded Gigaword as a corpus containing a large number of article-headline pairs for training an encoder-decoder model. Their work assumed a task setting where the first sentence of an article is a source text and its corresponding headline is a target text (summary). Since then, it has been a common practice to use the Gigaword dataset with this task setting and to measure the quality of generated headlines with ROUGE scores (Lin and Hovy, 2003) between system-generated and reference headlines.

Apparently, a summarization method is desirable to achieve a ROUGE score of 100, i.e., a system output is identical to the reference. However, this is an unrealistic goal for the task setting on the Gigaword dataset. The summarization task is underconstrained in that the importance of a piece of information highly depends on the expectations and prior knowledge of a reader (Kryściński et al., 2019). In addition, the Gigaword dataset (as well as other widely-used datasets) was noisy for summarization research because it was not created for the research objective but other professional activities (e.g., news production and distribution). Thus, the state-of-the-art method could only reach ROUGE-1 scores less than 40 on the dataset.

While a number of methods compete with each other for the underconstrained task on the noisy data, we have another concern about the *truthfulness* of generated summaries: whether all facts of a generated summary are mentioned in the source text. Unlike extractive summarization, abstractive summarization has no guarantee of truthfulness. This may result in a serious concern of practical applications of abstractive summarization when a generated summary includes fake facts that are not mentioned in the source document.

In this paper, we explore improving the truthfulness in abstractive summarization on two datasets, English Gigaword and JApanese MUlti-Length Headline Corpus (JAMUL) (Hitomi et al., 2019). In Section 2, we analyze headlines generated by the state-of-the-art encoder-decoder model and show that the model sometimes generates unexpected words. In order to estimate the truthfulness to the original text, we measure the recall-oriented ROUGE-1 scores between the source text and the generated headlines. This analysis reveals that a high ROUGE score between a reference and headline does not necessarily mean a high truthfulness to the source and that there is only a weak correlation between the two.

In Section 3, we conjecture that one of the reasons why the model sometimes exhibits such an untruthful behavior lies in untruthful article-headline pairs, which are used for training the model. In order to quantify the truthfulness of article-headline pairs, we consider the textual entailment of whether an article (source document) entails its headline. We will show that about 30–40% of source documents do not entail their headlines under the widely-used experimental settings. In other words, the current task setting is inappropriate for abstractive summarization. We release the annotations of textual entailment for both English Gigaword and JAMUL[1].

After confirming the untruthfulness of article-headline pairs in the datasets, we hypothesize that removing untruthful instances from the training data may remedy the problem of the untruthful behavior of the model. In Section 4, we build a binary classifier that predicts an entailment relation between an article and its headline and use the classifier to filter out untruthful instances in the training data. We train a model on the filtered supervision data in Section 5. Experimental results demonstrate that the filtering procedure shows no clear difference in ROUGE scores but remarkable improvements when we manually and automatically evaluate the truthfulness of the generated headlines. These results suggest the importance of evaluating truthfulness in addition to relevance.

2 Unexpected outputs

2.1 Examples of unexpected outputs

Although the current state-of-the-art method for abstractive summarization could only achieve a ROUGE-1 score of less than 40 on the Gigaword dataset, generated headlines actually look very fluent. This is probably because the encoder-decoder model acquired a strong language model from the vast amount of supervision data. However, some studies reported that the generated headlines often deviate from the content of the original document (Cao et al., 2018b; Kryściński et al., 2019). They addressed the problem where an abstractive model made mistakes in facts (e.g., tuples of subjects, predicates, and objects).

However, we also regularly see examples where the abstractive model generates unexpected words. This is true even for the state-of-the-art model.

Table 1 shows examples of unexpected outputs from UniLM (Dong et al., 2019), which shows the highest ROUGE scores[2] on English Gigaword. In the first example, the output includes "in November" whereas the input did not mention the exact month. In fact, this article was published in August 2009; however, the model probably guessed the month from the expression "this fall". The second example also exhibits a similar problem where the model incorrectly supplemented the news source "the Detroit News". The third and fourth examples are more problematic in that the generated headlines do not summarize the input sentences at all.

2.2 Estimating truthfulness

In order to quantify the problem of outputs that are untruthful to source documents, we measure the word overlap between the input and output of the UniLM model on the test set of English Gigaword (Rush et al., 2015). Here, we calculate the recall-oriented ROUGE-1 score[3], regarding an out-

[1] https://github.com/nlp-titech/headline-entailment

[2] UniLM model fine-tuned on Gigaword dataset achieved 38.90 ROUGE-1, 20.05 ROUGE-2, and 36.00 ROUGE-L scores as of November 22, 2019.

[3] We used SumEval: https://github.com/chakki-works/sumeval

#	Input (lead sentence)	Output (generated headline)
1	u.s. home resales posted the largest monthly increase in at least ## years last month as first-time buyers rushed to take advantage of a tax credit that expires this fall .	home sales rise #.# percent in november
2	seattle – for years , the standard treatment for patients with blood clots in veins deep in a limb has been blood thinners that stop the clots from getting bigger .	UNK drug may help treat UNK clots the detroit news
3	wigan moved to consolidate their premiership status tuesday by tying down one of the brightest stars of last season 's maiden top flight campaign .	english football league tables
4	never mind that she has dark blond hair and light blue eyes and the fairest of skin .	african-american girl is a UNK

Table 1: Examples of unexpected outputs generated by the state-of-the-art model. '#' stands for a digit mask. 'UNK' denotes an out-of-vocabulary word. The underlined parts indicate unexpected words.

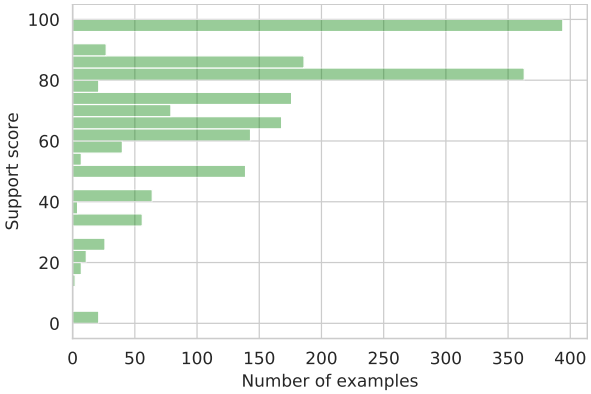

Figure 1: Histogram of support scores (recall-oriented ROUGE-1 scores between generated headlines and their source documents).

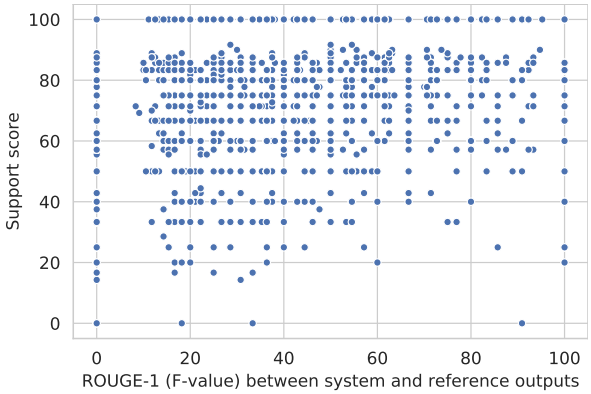

Figure 2: Scatter plots of ROUGE scores and support scores: X-axis presents ROUGE-1 score between system and reference headlines; and Y-axis presents support score (the same to Figure 1).

put (generated headline) as a *gold standard* and an input (source document) as a *target to be evaluated*[4]. Although this use of the ROUGE metric is unconventional, the intention here is to measure how many words in a generated headline originate from the input document. In other words, if all words in a generated headline are covered by its source document (truthful), the score is 100; if none of the words in a generated headline originate from its source document (untruthful), the score is 0. We call this ROUGE score *support score* hereafter to avoid naming conflicts with conventional ROUGE scores between system and reference summaries. We mention that we can find a similar method to the support score in several studies; for example, Zhang et al. (2018) measured the abstractiveness of an output. Our support score is roughly a reverse version of abstractiveness because the abstractiveness measures the number of words in an output that do not appear in the input.

Figure 1 reports the histogram of the support scores. A certain amount of instances receive relatively high support scores: 50.10% of the instances obtain scores larger than 80. At the same time, a non-negligible amount (9.14%) of instances have support scores less than 40. Note that the support scores present rough estimations of the truthfulness of the model; a lower score may imply that a headline includes paraphrased or shortened words from its source document. Having said that, Figure 1 indicates that the state-of-the-art model sometimes generates untruthful headlines.

Here, another interesting question comes into our mind: how do the widely-used benchmarking performance values (measured by ROUGE scores between system and reference headlines) reflect the

[4]We ignore instances whose source documents are less than ten characters long. The total number of instances after this treatment is 1,936.

truthfulness (measured by the support scores)? Figure 2 depicts the correlation between the two: the X-axis presents the ROUGE-1 score between system and reference headlines, and Y-axis presents support score. Unfortunately, we cannot observe a strong correlation between the two scores: Pearson's correlation coefficient between the two scores is 0.189, which suggests no correlation. This result supports that the conventional ROUGE scores tell us little about the truthfulness of generated summaries.

3 Are the task settings truthful?

3.1 Background of the datasets and settings

Why does a headline generation model exhibit untruthful behavior as we saw in the previous section? Before discussing the reason behind this, we need to understand how the datasets and task settings were established.

The Annotated English Gigaword corpus[5] is one of the most popular corpora in abstractive summarization research. Rush et al. (2015) converted this corpus into a dataset for abstractive summarization. They assumed the lead (first) sentence of an article as a source document and its corresponding headline as a target output. They did not explain the reason why they did not use a full-length article but only a lead sentence as a source document for headline generation. We infer that the reason for this treatment is that: a lead sentence provides a strong baseline for extractive summarization; their intention was to explore the capability of abstractive summarization from a lead sentence to a headline; using full text was time-consuming for encoder-decoder models.

Moreover, Rush et al. (2015) introduced some heuristics to remove some noisy instances. They discarded an instance if: (1) the source and target documents have no non-stop word in common; (2) the headline contains a byline or other extraneous editing marks; and (3) a headline includes a question mark or colon.

JApanese MUlti-Length Headline Corpus (JAMUL)[6] is a dataset specially designed for evaluating summarization methods. JAMUL consists of 1,524 Japanese full-text articles and their print headlines (used for newspapers). Although JAMUL is distributed for free of charge, JAMUL alone is insufficient for training an encoder-decoder model. Hitomi et al. (2019) also released Japanese News Corpus (JNC), which is a large-scale dataset consisting of 1,831,812 pairs of newspaper articles and their print headlines. JNC includes only the first three sentences of each article[7].

Table 2 summarizes the datasets and task settings. As we can see from the rows of Rush et al. (2015) and JNC, these task settings do not use full-text articles but only lead (6.6% of words in full articles, Gigaword) and lead three sentences (25.9% of words in full articles, JNC) as source documents for abstractive summarization. Hence, we hypothesize that the source documents under these task settings contain insufficient information for generating headlines. In other words, headline generation models might be faced with supervision data where headlines cannot be generated from source documents and learned to be untruthful, i.e., producing pieces of information that are not mentioned in source documents.

3.2 Truthfulness of the datasets and settings measured by textual entailment

This section explores the hypothesis: *do source documents include sufficient information to produce headlines?* We examine this hypothesis by considering textual entailment between a source document and its headline. More specifically, we would like to know whether a source document entails its headline, i.e., whether we can infer that a headline is true based on the information in the source document.

We asked three human subjects to judge entailment relations for 1,000 pairs of source documents and headlines of each dataset. We randomly selected 1,000 pairs from the test set of the English Gigaword dataset and 1,000 pairs from JAMUL. The labels include *entail*, *non-entail*, and *other* (see Appendix for the definition of the labels and the treatment).

Table 4 reports the ratio of document-headline pairs for which two or three human subjects voted 'yes' for the entailment relation (*entail*). Only 70.3% of lead-headline pairs in the Gigaword dataset hold the entailment relation. For reference, we did the same analysis by using full-text articles as source documents and found that the ratio

[5] https://catalog.ldc.upenn.edu/LDC2012T21

[6] https://cl.asahi.com/api_data/jnc-jamul-en.html

[7] This is because the price of the dataset would be much higher if it included full-text articles.

data	# docs	# words	# sent / doc	# words / doc	# words / headline
English Gigaword	8.6 M	77 M 4 B	20.3	477.6	8.9
Rush et al. (2015)	3.8 M	31 M 119 M	1	31.3	8.3
JAMUL	1.5 k	23 k 547 k	11.7	359.2	15.3
JNC	1.8 M	26 M 171 M	3	93.2	14.2

Table 2: The statistics of datasets and task settings. The column "# words" presents two values for each row: a top value is the total number of words in the headline; and the bottom value is the total number of words in the article. The second row of each group (Rush et al. (2015) and JNC) corresponds to the setting of training data. The columns "# sent / doc", "# words / doc", and "# words / headline" denote the average number of sentences per source document, words per source document, and words per headline, respectively.

#	Source document (text)	Headline (hypothesis)	Entail
1	France *hopes to secure the contract for the supply of Agosta-class submarines* to the Malaysian navy...	France *keen to sell submarines* to Malaysian navy	Y
2	*69,700 local people to work*	*70,000 employees*	Y
3	British boxing promoter Frank Warren on Tuesday announced the signing of three world title contenders.	Three foreign boxers join British stable	N
4	Lazio and Roma will be playing for more than local bragging rights when they meet...	Football : Italian Serie A table	N

Table 3: Example of entailment labels between source document (text) and headline (hypothesis). An italic part presents a paraphrase, and an underlined part presents a deviation.

Dataset	Lead-1	Lead-3	Full
Gigaword	70.3%	N/A	92.8%
JAMUL	N/A	61.4%	94.2%

Table 4: Ratio of document-headline pairs where the source documents entail their headlines.

rises to 92.8%. Similarly, only 61.4% of lead three sentences (lead-3) and headline pairs in JAMUL hold the entailment relation. When using full-text articles, the entailment ratio rises to 94.2%. These results support our hypothesis that source documents contain insufficient information under the current task settings.

4 Improving the truthfulness of data

Based on the analysis in the previous section, we can consider two strategies to improve the task setting: using full-text articles as source documents instead of leading sentences; and removing non-entailment instances from the dataset. Although the former strategy reduces the ratio of non-entailment pair to 7.2% (English Gigaword) and 5.8% (JAMUL), we must consider the trade-off: the use of full-text articles increases the cost for training, and may decrease the quality of headlines because of longer inputs to encoder-decoder models. Furthermore, JNC does not provide full-text articles but only lead three sentences. Therefore, we take the latter strategy, removing non-entailment pairs from the supervision data for headline generation.

4.1 Recognizing textual entailment

In order to find non-entailment pairs in the dataset, we build a binary classifier that judges whether a source document entails its headline or not. Recently, pretrained language models such as BERT (Devlin et al., 2019) show remarkable advances in the task of recognizing textual entailment (RTE)[8]. Thus, we fine-tune pretrained models on the supervision data for entailment relation between source documents and their headlines.

For English Gigaword dataset, we use the pretrained RoBERTa large (Liu et al., 2019) fine-tuned on Multi-Genre Natural Language Inference (MultiNLI) (Williams et al., 2018). We further fine-

[8] https://gluebenchmark.com/leaderboard

tuned the model on the supervision data of the lead-headline pairs with entailment labels (acquired in Section 3). Here, the supervision data include lead-headline pairs where two or three human subjects labeled either *entail* or *non-entail*; other pairs were excluded from the supervision data. In this way, we obtained a binary classifier for entailment relation of 91.7% accuracy on a hold-out evaluation (761 training and 179 test instances) after running 10 epoch of fine-tuning on the RoBERTa model.

For JNC, we use the pretrained BERT model for Japanese text (Kikuta, 2019). However, no large-scale Japanese corpus for semantic inference (counterpart to MultiNLI) is available. Thus, we created supervision data for entailment relation between lead three sentences and headlines (*lead3-headline*, hereafter) on JNC. We extracted 12,000 lead3-headline pairs from JNC, and collected entailment labels using crowdsourcing. Each pair had five entailment labels assigned by five crowd workers. We used lead3-headline pairs where four or five crowd workers labeled either *entail* or *non-entail*; other pairs were unused in the supervision data. The entailment classifier fine-tuned on the supervision data achieved 83.9% accuracy on a hold-out evaluation with 5,033 training and 1,678 test instances.

Applying the entailment classifiers to the training and development sets of English Gigaword dataset and JNC, we removed instances of non-entailment pairs judged by the classifiers. Eventually, we obtained 2,695,325 instances (71% of the original training instances) on the English Gigaword dataset and 841,640 instances (49% of the original training instances) on JNC.

5 Improving the truthfulness of models

In this section, we examine whether the supervision data built in the previous section reduces untruthful headlines.

5.1 Headline generation models

We use fairseq[9] (Ott et al., 2019) as an implementation of the Transformer architecture (Vaswani et al., 2017) throughout the experiments. Hyper-parameter configurations are: 6 layers both in the encoder and decoder; 8 attention heads; the dimension of hidden states is 512; the dimension of hidden states of the feed forward network is 2048; the smoothing rate, dropout rate, and label smoothing were set to 0.1; Adam optimizer with $\beta = 0.98$, the learning rate of 0.0005, and 4,000 warm-up steps.

We train the Transformer models on the supervision data with and without non-entailment instances. Because removing non-entailment instances decreases the number of training instances, we also apply the self-training strategy (Murao et al., 2019) to obtain the same amount of training instances to the full supervision data. More specifically, we generated headlines for the source documents discarded in Section 4.1, and added pairs of source documents and generated headlines as pseudo supervision data. The experiments compare models trained on the full supervision data (*full*), the one filtered by the entailment classifier (*filtered*), and the one filtered but augmented by the self-training (*filtered+pseudo*).

5.2 Data preparation

The experiments use the same data split of training (3.8M instances), development (390k instances), and test (380k instances) sets to Rush et al. (2015). In this study, we used 10,000 instances for evaluation that were sampled from the test set and unused in the analysis in Section 3. We do not apply any replace operations for the English Gigaword dataset: digit masking, rare word to UNK, and lower-casing. The dataset is tokenized by WordPiece (Wu et al., 2016) with the same vocabulary used in UniLM.

Splitting JNC into 1.7M training and 3k development instances, we evaluate the model on the JAMUL dataset. We use SentencePiece[10] (Kudo and Richardson, 2018) for tokenization.

5.3 Evaluation protocol

We evaluate the quality of generated headlines by using full-length F1 ROUGE scores[11], following the previous work. However, Kryściński et al. (2019) reported that ROUGE scores between system and reference summaries had only a weak correlation with human judgments. Furthermore, we would like to confirm whether the filtering strategy can improve the truthfulness of the model. Therefore, we also report the support score, the ratio of entailment relation between source documents and generated headlines measured by the entailment classifiers (explained in Section 4.1), and human evaluation about the truthfulness.

[9] https://github.com/pytorch/fairseq

[10] https://github.com/google/sentencepiece

[11] ROUGE scores were computed by SumEval.We used MeCab (Kudo et al., 2004) for Japanese tokenization.

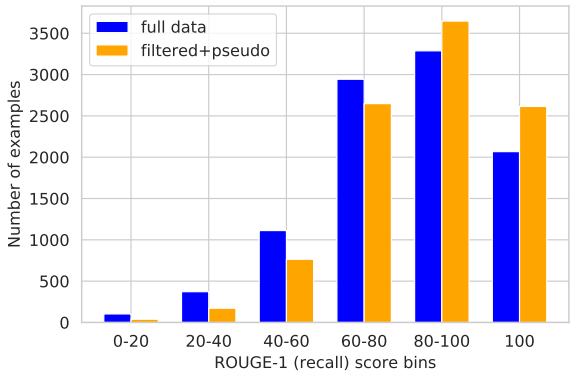

Figure 3: The distribution of the support scores on the English Gigaword dataset.

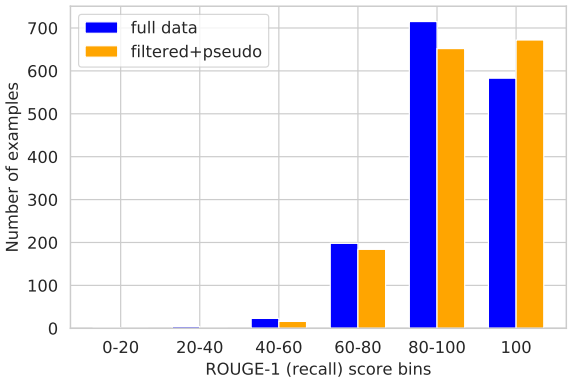

Figure 4: The distribution of the support scores on JAMUL.

5.4 Results

Table 5 shows the main results. The baseline model with full training data obtained 35.80 ROUGE-1 score on the English Gigaword dataset and 48.08 ROUGE-1 score on JAMUL. The entailment filter lowered ROUGE scores on both of the datasets probably because of the smaller number of training instances, but the self-training strategy improved ROUGE scores on the Gigaword dataset, outperforming the baseline model.

In contrast, the self-training strategy could not show an improvement for ROUGE scores on JAMUL. Although it is difficult to find the exact cause of this result, we suspect that the filtering step reduced the training instances too much (0.8M instances) for the self-training method to be effective. Another possibility is that the writing style of articles of non-entailment pairs in JNC/JAMUL is so distant that the self-training method generated headlines that are too different from reference ones.

The column "Sup" presents the support score computed by the recall-oriented ROUGE-1 between source documents and generated headlines (explained in Section 2.2). The table indicates that the filtering and self-training strategies obtain higher support scores than the baseline. Figures 3 and 4 depict histograms of the support scores for the baseline and filtering+pseudo settings on Gigaword and JAMUL, respectively. We could confirm that the filtering+pseudo strategy increased the number of headlines with high support scores.

The column "Entail" shows the entailment ratio measured by the entailment classifier. Again, the filtering+pseudo strategy obtained the highest entailment ratio on both the Gigaword dataset and JAMUL. Although this result may be interpreted as natural because we selected training instances based on the same entailment classifier, it is interesting to see that we can control the entailment ratio without changing the model.

In order to examine whether the filtering strategy can deliver noticeable improvements for human readers, we asked a human subject to judge the truthfulness of the headlines generated by the baseline setting and filtering+pseudo strategy. Presented with both a source document and a headline generated by the model, the human subject judged whether the headline was *truthful*, *untruthful*, or *incomprehensible*. We conduct this evaluation for 109 instances randomly sampled from the test sets of Gigaword and JAMUL.

The "Truthful" column in Table 5 reports the ratio of truthful headlines. Consistently with the entailment ratio, we could confirm that the filtering+pseudo strategy generated truthful headlines more than the baseline setting on both of the datasets. During the human evaluation, one instance in both full and filtered+pseudo settings from the Gigaword dataset judged as incomprehensible.

5.5 Discussion

To sum up the results, improving the truthfulness of the supervision data does help improving the truthfulness of generated headlines. We could confirm the improvements from the support scores, entailment ratio, and human judgments. However, the ROUGE scores between system and reference headlines did not indicate a clear difference.

The ROUGE metric was proposed to measure the *relevance* of a summary when extractive summarization was the central approach (in the early 2000s). Obviously, the truthfulness of summaries

Dataset	Training data (amount)		R-1	R-2	R-L	Sup	Entail	Truthful
Gigaword	Full	(3.8 M)	35.80	17.63	33.69	75.38	85.78%	77.06%
	Filtered	(2.7 M)	35.24	17.29	33.14	77.61	91.50%	—
	Filtered+pseudo	(3.8 M)	**35.85**	**17.94**	**33.72**	**79.91**	**93.56%**	**85.32%**
JAMUL	Full	(1.7 M)	**48.08**	**22.21**	**40.02**	89.10	90.29%	89.91%
	Filtered	(0.8 M)	46.08	20.81	38.07	90.14	95.67%	—
	Filtered+pseudo	(1.7 M)	45.62	20.55	38.10	**90.65**	**96.26%**	**92.66%**

Table 5: Results on the test set. We used F1 full-length ROUGE score: R-1 (ROUGE-1), R-2 (ROUGE-2), and R-L (ROUGE-L). "Sup" denotes support score. "Entail" presents the percentage of outputs to which the entailment classifier predicts the entailment relation (built in Section 4.1). "Truthful" show the percentage of outputs to which a human subject judged as truthful headlines.

is out of the scope of ROUGE. The experimental results in this paper suggest that we should consider both *relevance* and *truthfulness* when evaluating the quality of abstractive summarization.

6 Related Work

Rush et al. (2015) first applied the neural sequence-to-sequence (seq2seq) architecture (Sutskever et al., 2014; Bahdanau et al., 2015) to abstractive summarization. They obtained a dataset for abstractive summarization from the English Gigaword (Graff and Cieri, 2003; Napoles et al., 2012). After this work, a large number of studies followed the task setting (Takase et al., 2016; Zhou et al., 2017; Cao et al., 2018a; Song et al., 2019; Wang et al., 2019).

Some researchers pointed out that abstractive summarization models based on seq2seq sometimes generate summaries with inaccurate facts. Cao et al. (2018b) reported that 30% of the summaries generated by a seq2seq model include different facts from source articles. In addition, Kryściński et al. (2019) reported that ROUGE scores have only a weak correlation with human judgments in abstractive summarization and that the current evaluation protocol is inappropriate for factual consistency.

Several studies approach the problem of inconsistency between input and output by improving the model architecture or learning method. Cao et al. (2018b) applied an information extraction tool to extract tuples of subject, predicate, and object from source documents and utilized them as an additional input to the model. Pasunuru and Bansal (2018) incorporated an entailment classifier as a reward in reinforcement learning. Guo et al. (2018) presented a multi-task learning method between summarization and entailment generation where hypotheses entailed by a given document (as a premise) are generated. Li et al. (2018) introduced an entailment-aware encoder-decoder model to ensure the correctness of the summary. Kiyono et al. (2018) reduced incorrect generations by modeling token-wise correspondences between input and output. Falke et al. (2019) proposed a re-ranking method of beam search based on factual correctness from a classifier of textual entailment.

As another direction, Kryscinski et al. (2019) evaluated the factual consistency of a source document and the generated summary with a weakly-supervised model.

A few studies raised concerns about the data set and task setting. Tan et al. (2017) argued that lead sentences do not provide an adequate source for the headline generation task. The researchers reported that making use of multiple summaries as well as the lead sentence of an articles improved the performance of headline generation on the New York Times corpus. In contrast, our paper is the first to analyze the truthfulness of existing datasets and generated headlines, provide a remedy to the supervision data, and demonstrate the importance of truthfulness in headline generation.

7 Conclusion and future work

In this paper, we showed that the current headline generation model yields unexpected words. We conjectured that one of the reasons lies in the defect in the task setting and data set, where generating a headline from the source document is impossible because of the insufficiency of the source information. We presented an approach for removing from the supervision data headlines that are not entailed by their source documents. Experimental results demonstrated that the headline generation model trained on filtered supervision data showed no clear difference in ROUGE scores but remarkable im-

provements in automatic and manual evaluations of the truthfulness of the generated headlines. We also presented the importance of evaluating truthfulness in abstractive summarization.

In the future, we explore a more sophisticated method to improve the relevance and truthfulness of generated headlines, for example, removing only deviated spans in untruthful headlines rather than removing untruthful headlines entirely from the supervision data. Other directions include an extensive evaluation of relevance and truthfulness of abstractive summarization and an establishment of an automatic evaluation metric for truthfulness.

Moreover, it will be also interesting to see whether the same issue occurs in other related tasks such as data-to-text generation. We believe that the concern raised in this paper is beneficial to other tasks.

Acknowledgments

The research results have been achieved by "Research and Development of Deep Learning Technology for Advanced Multilingual Speech Translation", the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018a. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 152–161.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018b. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 4784–4791.

James Clarke and Mirella Lapata. 2008. Global inference for sentence compression an integer linear programming approach. *Journal of Artificial Intelligence Research*, 31(1):399–429.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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 (NAACL-HLT)*, pages 4171–4186.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2214–2220.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization*, pages 40–48.

David Graff and Christopher Cieri. 2003. English Gigaword (LDC2003T05). https://catalog.ldc.upenn.edu/LDC2003T05, Linguistic Data Consortium, Philadelphia.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 687–697.

Yuta Hitomi, Yuya Taguchi, Hideaki Tamori, Ko Kikuta, Jiro Nishitoba, Naoaki Okazaki, Kentaro Inui, and Manabu Okumura. 2019. A large-scale multi-length headline corpus for improving length-constrained headline generation model evaluation. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG)*, pages 436–441.

Yohei Kikuta. 2019. BERT pretrained model trained on Japanese Wikipedia articles. https://github.com/yoheikikuta/bert-japanese.

Shun Kiyono, Sho Takase, Jun Suzuki, Naoaki Okazaki, Kentaro Inui, and Masaaki Nagata. 2018. Reducing odd generation from neural headline generation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC)*.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization - step one: Sentence compression. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth*

Conference on Innovative Applications of Artificial Intelligence (IAAI), pages 703–710.

Wojciech Kryściński, 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. 2019. Evaluating the factual consistency of abstractive text summarization. *ArXiv*, abs/1910.12840.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 66–71.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 230–237.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics (Coling)*, pages 1430–1441.

Chin-Yew Lin and Eduard Hovy. 2003. 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 (HLT-NAACL)*, pages 150–157.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 510–520.

Yinhan Liu, Myle Ott, Naman 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.

Inderjeet Mani. 2001. *Automatic Summarization*. John Benjamins Publishing Company.

Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 170–173.

Kazuma Murao, Shintaro Takemae, Hayato Kobayashi, Taichi Yatsuka, Masaki Noguchi, Hitoshi Nishikawa, and Takenobu Tokunaga. 2019. Neural headline generation with self-training. In *Proceedings of Computation+Journalism Symposium 2019*.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, pages 95–100.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Ramakanth Pasunuru and Mohit Bansal. 2018. 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 (NAACL-HLT)*, pages 646–653.

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 (EMNLP)*, pages 379–389.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning (ICML)*, pages 5926–5936.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, pages 3104–3112.

Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural headline generation on abstract meaning representation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1054–1059.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4109–4115.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, pages 5998–6008.

Kai Wang, Xiaojun Quan, and Rui Wang. 2019. BiSET: Bi-directional selective encoding with template for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2153–2162.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1112–1122.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Fangfang Zhang, Jin-ge Yao, and Rui Yan. 2018. On the abstractiveness of neural document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 785–790.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1095–1104.

Entail

- All facts of the headline are covered by those of the article.
- If the headline includes an expression that do not appear in the article, but if the fact mentioned by the expression can be derived from the article, judge the pair as "Entail".

Non-entail

- The statement of the headline conflicts with the article.
- The headline mentions facts that cannot be confirmed by the article.

Incomprehensible

- Impossible to judge because the article or headline is unreadable. If the headline is not grammatically complete but correct as the headline style, please try to judge either entail or non-entail.
- Other problems such as garbled characters.

Figure 5: Guideline for entailment labeling

A Guideline for entailment labeling

Figure 5 presents a guideline for the entailment labeling task in Section 3. Given a pair of an article and headline, a crowd worker is expected to judge whether the article entails the headline, and label the pair with either of the labels shown in this figure.

B Examples

Figure 6 shows some examples of the generated headlines from the models described in Section 5. In the first example, the baseline model added "in Kashmir" in the headline, but this is incorrect. The correct location is in Southern Egypt, which was mentioned in the reference headline. The filtered+pseudo model generates a safe headline. The second headline generated by the baseline includes the verb 'begin' although the report was written two years ago. The baseline model added "dollar lower against yen" in the headline. There is a correlation indeed that dollar is lower against yen when Tokyo stocks rise, but we cannot confirm the fact from the source document.

Source:
Suspected Muslim militants shot and killed five men who had formed a civilian patrol group to counter the radicals in their village , police officials said Monday .
Full (baseline):
Suspected Militants Kill Five in Kashmir
Filtered+pseudo:
Suspected Muslim Militants Kill Five

Source:
Divers searched the Mississippi River for bodies still trapped beneath the twisted debris of a collapsed freeway bridge Thursday , as finger - pointing began over a federal report two years ago that found the bridge was ' ' structurally deficient .
Full (baseline):
FEDERAL REPORT BEGINS IN MISSISSIPPI
Filtered+pseudo:
Divers Search Mississippi River for Bodies

Source:
Tokyo stocks rose Tuesday as investors snapped up domestic demand - related issues due to receding jitters among investors over last week ' s plunge .
Full (baseline):
Tokyo stocks rise , dollar lower against yen
Filtered+pseudo:
Tokyo stocks end higher

Figure 6: Examples of the improved headlines.