Answers Unite! Unsupervised Metrics for Reinforced Summarization Models

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano

Answers Unite!
Unsupervised Metrics for Reinforced Summarization Models

Thomas Scialom[⋆‡], **Sylvain Lamprier**[‡], **Benjamin Piwowarski**[◇‡], **Jacopo Staiano**[⋆]
[◇] CNRS, France
[‡] Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
[⋆] reciTAL, Paris, France
{thomas,jacopo}@recital.ai
{sylvain.lamprier,benjamin.piwowarski}@lip6.fr

Abstract

Abstractive summarization approaches based on Reinforcement Learning (RL) have recently been proposed to overcome classical likelihood maximization. RL enables to consider complex, possibly non-differentiable, metrics that globally assess the quality and relevance of the generated outputs. ROUGE, the most used summarization metric, is known to suffer from bias towards lexical similarity as well as from suboptimal accounting for fluency and readability of the generated abstracts. We thus explore and propose alternative evaluation measures: the reported human-evaluation analysis shows that the proposed metrics, based on Question Answering, favorably compares to ROUGE – with the additional property of not requiring reference summaries. Training a RL-based model on these metrics leads to improvements (both in terms of human or automated metrics) over current approaches that use ROUGE as a reward.

1 Introduction

Summarization systems aim at generating relevant and informative summaries given a variable-length text as input. They can be roughly divided under two main categories, those adopting an *extractive* approach, *i.e.* identifying the most informative pieces from the input text and concatenating them to form the output summary; and those producing *abstractive* summaries, *i.e.* generating an output text whose tokens are not necessarily present in the input text.

While closer to human summarization, abstractive summarization is a much harder task and the need for faithful evaluation metrics is crucial to measure and drive the progress of such systems. The standard for evaluation of summarization systems is ROUGE (Lin, 2004): this metric can be considered as an adaptation of BLEU (Papineni et al., 2002), a scoring method for evaluation of machine translation systems; both based on *n*-gram co-occurrences, the latter favors precision while the former emphasizes recall.

Recent research works (Paulus et al., 2017; Pasunuru and Bansal, 2018; Arumae and Liu, 2019) have proposed to use evaluation metrics – and ROUGE in particular – to learn the model parameters through Reinforcement Learning (RL) techniques. This makes the choice of a good evaluation metric even more important. Unfortunately, ROUGE is known to incur several problems: in particular, its poor accounting for fluency and readability of the generated abstracts, as well as its bias towards lexical similarity (Ng and Abrecht, 2015). To emphasize the latter point, since ROUGE evaluates a summary against given human references, summarization models incur the risk of being unfairly penalized: a high quality summary might still have very few tokens/*n*-grams in common with the reference it is evaluated against.

In this work, we propose to overcome *n*-gram matching based metrics, such as ROUGE, by developing metrics which are better predictors of the quality of summaries. The contributions of this paper can be summarized as follows:

- Extending recent works (Eyal et al., 2019; Chen et al., 2018), we introduce new metrics, based on Question Answering, that do not require human annotations.
- We report a quantitative comparison of various summarization metrics, based on correlations with human assessments.
- We leverage the accuracy of the proposed metrics in several reinforcement learning schemes for summarization, including two unsupervised settings: *in-domain* (raw texts from the target documents) and *out-of-domain* (raw texts from another document collection).

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3246–3256, Hong Kong, China, November 3–7, 2019.

- Besides a quantitative evaluation of the generated summarizes, we qualitatively evaluate the performances of the different approaches through human assessment.

Our main results can be summarized as follows:

1. We show that fitting human judgments from carefully chosen measures allows one to successfully train a reinforcement learning-based model, improving over the state-of-the-art (in terms of ROUGE and human assessments).

2. we show that dropping the requirement for human-generated reference summaries, as enabled by the proposed metrics, allows to leverage texts in a self-supervised manner and brings clear benefits in terms of performance.

Section 2 introduces the metrics. Section 3 reviews related summarization systems and presents our proposed approaches. Section 4 presents our experimental results and discussions.

2 Evaluation Metrics

This section first describes our selection of existing summarization metrics and introduces our proposals. Then, we quantitatively compare them for abstractive summarization. For a comprehensive list of evaluation metrics, we refer the reader to Liu et al. (2016).

2.1 *n*-grams-based metrics

TextRank Automated summarization started with the development of extractive text summarization models. Many unsupervised models, that aim at computing a score between a sentence and document(s) were developed – the score attempting to reflect whether the sentence should be selected for building a summary (Nenkova, 2011). Such scores can thus be used as a proxy of the summary quality. We chose TextRank (Mihalcea and Tarau, 2004) – an extractive non-parametric summarization system inspired by PageRank (Page et al., 1999) – since it is well performing for extractive tasks and could be easily adapted for our needs. The algorithm builds a graph of sentences within a text based on their co-occurrences. Then, it assigns an importance score for each sentence based on a random walk on the resulting graph. The most important elements of the graph are considered as the ones that best describe the text. As a derivative usage, we propose to consider these importance scores to assess the quality of abstractive summaries in our study. This metric is referred to as TextRank in the following.

ROUGE Arguably the most popular metric for summarization at the moment, it provides a set of measures to compare automatically generated texts against one or more references (Lin, 2004). In particular, ROUGE-N is based on the count of overlapping *n*-grams, while ROUGE-L accounts for the longest common sub-sequence between the candidate and its corresponding reference(s).

Novelty As noted by See et al. (2017), abstractive summarization models do not produce novel *n*-grams as often as the reference summaries. Thus, to favor the generation of unseen words and produce more abstractive summaries, Kryściński et al. (2018) integrated *novelty* as a reward for reinforcement learning. It is defined as the fraction of unique *n*-grams in the summary that are novel, normalized by the length ratio of the generated and reference summaries.

2.2 Beyond n-grams

2.2.1 Language Modeling

We investigate the use of language models as an evaluation metric. (ShafieiBavani et al., 2018) proposed to exploit word embeddings to train a model able to rate the generated summaries. Following neural language models (LM), we propose to consider the perplexity of the generated summary according to the BERT LM (Devlin et al., 2019), which demonstrated state of the art results in many NLP tasks. For our experiments, we used the publicly available pre-trained English "base" model.

2.2.2 Question-Answering based Metrics

Question-Answering is related to summarization: the first work in this direction (Wu et al., 2002) introduced the notion of Answer-Focused Summarization, where answers to relevant questions on the source text are used to build the corresponding summary. Based on the intuition that a good-quality summary should provide the answers to the most relevant questions on a given input text, several works have proposed to adapt Question Answering (QA) for summary quality evaluation.

In that vein, (Pasunuru and Bansal, 2018) proposed to measure if answers contain the most *salient* tokens. Closer to our work, (Eyal et al.,

2019) proposed *APES*, a novel metric for evaluating summarization, based on the hypothesis that the quality of a generated summary is linked to the number of questions (from a set of relevant ones) that can be answered by reading it. In their proposed setup, two components are thus needed: (a) a set of relevant questions for each source document; and (b) a QA system. For each summary to assess, questions are successively generated from a reference summary, by masking each of the named entities present in this reference, following the methodology described in (Hermann et al., 2015). This results in as many triplets $(input, question, answer)$ as named entities present in the reference summary, where *input* denotes the summary to assess, *question* refers to the sentence containing the masked entity and *answer* refers to this masked entity to retrieve. Thus, for each summary to assess, metrics can be derived from the ability of the QA system to retrieve the correct answers from each of the associated triplets.

F1 score For each triplet, an F1 score is computed according to the responses retrieved by the considered QA system. This score, commonly used for QA evaluation (Rajpurkar et al., 2016), measures the average overlap between predictions and ground truth answers. For each summary to assess, the is the average of the F1 score computed over each triplet. In the following, we denote this metric as QA_{fscore}*(sup)*.

QA confidence Complementary to the F1 score, we propose to also consider the confidence of the QA system for its retrieved answer. This corresponds, for each triplet, to the probability of the true answer according to the QA model. Confidence scores are averaged for each summary to assess over its associated triplets. In the following, we denote this metric as QA_{conf}*(sup)*.

Besides considering the simple presence of the expected answers in the generated summary, QA-based metrics also account to some extent for readability. They indeed require that the considered QA system, trained on natural language, be able to find the answer in the input to assess, despite the variability of the generated texts.

Extension to the unsupervised setting While being a useful complement to ROUGE, the two QA-based metrics described above still need human-generated summaries. In this paper, we investigate and propose extending the previously described QA-based approach in an unsupervised setting.

	Readability	**Relevance**
Readability	1.0	0.77 **
Relevance	0.77 **	1.0
ROUGE-1 (sup)	0.14 *	0.18 **
ROUGE-2 (sup)	0.12 *	0.18 **
ROUGE-L (sup)	0.13 *	0.18 **
TextRank (unsup)	0.14 *	0.13 **
Novelty (unsup)	-0.13 *	-0.1 *
Bert LM (unsup)	0.21 **	0.08 *
QA_{fscore} (sup)	0.14 *	0.19 **
QA_{conf} (sup)	0.19 **	0.23 **
QA_{fscore} (unsup)	0.08	0.2 **
QA_{conf} (unsup)	0.33 **	0.31 **

Table 1: Spearman's ρ for the different metrics w.r.t. Readability and Relevance (*: $p < .05$, **: $p < .005$).

With this aim, we extended the above metrics at *the document level* (*i.e.*, questions and answers are generated from the source article text rather than from the reference summary), dispensing of the need for human-generated reference summaries. Thus, in line with the *APES* approach described above, we propose two unsupervised QA-based metrics, to which we refer to as QA_{fscore}*(unsup)* and QA_{conf}*(unsup)*. Accounting for both quality and informativeness of a generated summary, those metrics have the appealing property of not requiring reference summaries.

2.3 Quantitative Analysis

We exploit human judgments obtained for 3 types of automatically generated summaries by Paulus et al. (2017) on 100 samples of the CNN/Daily Mail summarization dataset (see detail in section 4.1), in terms of readability (how well written the summary is) and relevance (how well does the summary capture the important parts of the article). The summaries are generated by the three different systems proposed in the original work. Those samples have been scored, via Amazon Mechanical Turk, for Readability and Relevance (scores from 1 to 10 for both metrics).

In Table 1, we report Spearman's rank correlations on this data, where we compare summaries rankings obtained according to the assessed metrics. Scores render the ability of the various metrics to reproduce human preferences (in terms

of readability and relevance). First, we observe that readability and relevance are naturally intertwined: intuitively, an unreadable summary will bear very little information, one of the facts that explains the high correlation between *readability* and *relevance*.

From this correlation analysis against human judgments, we observe that, as expected, the Language Model metric captures *readability* better than ROUGE, while falling short on *relevance*.

On the other hand, the results obtained using the proposed QA-based metrics indicate their potential benefits especially under the *unsupervised* setting, with QA_{conf} and QA_{fscore} capturing *readability* and *relevance* better than all the others reported metrics, including ROUGE. We thus conclude that the proposed metrics, which favorably correlate with *readability* and *relevance* under human evaluation, are worth of a deeper experimental investigation: in the following sections we provide a thorough evaluation of their contributions as Reinforcement Learning rewards signals.

2.4 Learned Metric

Finally, we also leverage the qualitative data obtained by Paulus et al. (2017) – which compounds to 50 samples evaluated by annotators in terms of *readability* and *relevance* – to *learn* an aggregate metric for evaluation. We use a Ridge regression (with a regularization $\lambda = 1$) to learn to predict the geometric mean of readability and relevance from the metrics defined above. The geometric means was chosen since we want the generated summary to be both readable and relevant.

We randomly sampled 50% of the data to fit the linear model with various subsets of our base metrics. Then, we measured the correlation w.r.t. the expected geometric mean on the remaining 50% data. We performed this procedure 1000 times. Our experiments show that the best performing set of metrics consists of ROUGE-L in conjunction with QA_{conf} and QA_{fscore}, both computed at an article-level, and hence unsupervised.

This learned metric is thus defined as (with *unsup* versions of QA-based scores):

$$\alpha ROUGE_L + \beta QA_{conf} + \delta QA_{fscore} \quad (1)$$

with $\alpha = 0.8576$, $\beta = 2.274$ and $\delta = 0.6413$. We leverage this learned metric in our RL-based summarization model, as described below.

2.5 Implementation details

As QA system we use the BERT "base" pre-trained model (Devlin et al., 2019), finetuned on the SQuAD dataset (Rajpurkar et al., 2016) using the recommended parameters for the task[1]. This differs from the approach adopted by (Eyal et al., 2019) who trained their QA model on CNN-DM (the same data used for the summarization task).

3 Summarization Models

Abstractive summarization systems were originally designed as a post-processing of an extractive system – by compressing sentences (Nenkova, 2011). A lot of activity takes place nowadays in designing neural networks sequence to sequence architectures (Sutskever et al., 2014), which allow to consider the problem as a whole rather than a two-step process. Such models reached state-of-the-art performance. To tackle the summarization, which deals with a long text and possibly out-of-vocabulary tokens, See et al. (2017) proposed to leverage an attention over the input (Bahdanau et al., 2014), as well as a copy mechanism (Vinyals et al., 2015).

One problem of sequence-to-sequence models is that they tend to repeat text in the output. To deal with this problem, (See et al., 2017) use a *coverage mechanism*, and Paulus et al. (2017) introduced *Intra-Decoder Attention* with the same goal of avoiding duplicate information within the output sequences.

More recently, the model proposed by See et al. (2017) was further extended (Gehrmann et al., 2018), with the addition of an attention mask during inference: a pre-trained sequence tagger trained to select which input tokens should be copied and used to filter the copy mechanism. Such a filter, called *Bottom-Up Copy Attention*, was shown to help prevent copying from the source text sequences that are too long, hence resulting in more abstractive summaries. On the CNN/Daily Mail dataset, (Gehrmann et al., 2018) found this two-step process to yield significant improvements in terms of ROUGE – resulting in the current state-of-the-art system. We base our experiments on this model.

The differentiable loss function commonly used for training summarization models, negative log-likelihood, has several known limitations. Among

[1] https://github.com/huggingface/pytorch-transformers.

those, exposure bias and failure to cope with the large number of potentially valid summaries.

To overcome this, approaches based on reinforcement learning have recently been proposed, allowing the models to learn via reward signals. Ranzato et al. (2015) used the REINFORCE algorithm (Williams, 1992) to train RNNs for several generation tasks, showing improvements over previous supervised approaches. Narayan et al. (2018) used such an approach in an extractive summarization setting, learning to select the most relevant sentences within the input text in order to construct its summary. (Paulus et al., 2017) combined supervised and reinforcement learning, demonstrating improvements over competing approaches both in terms of ROUGE and on human evaluation. However, the main limit of these works is that they rely on standard summarization metrics which are known to be biased.

Finally, closer to our work, Arumae and Liu (2019) proposed to use question-answering rewards to learn an *extractive* summarization model in a reinforcement learning setup. Compared to what we propose, their system is extractive, and relies on hand-written summaries.

3.1 Mixed Training Objectives

In our experiments, we follow the reinforcement learning scheme described below. The main difference with previous works is our reward function, which was based on our study of metrics (section 2). We consider a mixed loss L_{ml+rl} combining supervised and reinforcement learning schemes:

$$L_{ml+rl} = \gamma L_{rl} + (1-\gamma) L_{ml} \tag{2}$$

where we define the reinforcement loss L_{rl} and the maximum likelihood L_{ml} in the following paragraphs.

Maximum Likelihood Under a supervised training setup, the teacher forcing algorithm (Williams and Zipser, 1989) can be applied, and corresponds to maximizing the likelihood (ML) or equivalently to minimizing the negative log likelihood (NLL) loss defined as:

$$L_{ml} = -\sum_{t=0}^{m} \log p(y_t^* | y_0^*, ..., y_{t-1}^*, X) \tag{3}$$

where $X = [x_1, ..., x_n]$ is the input text of n tokens and $Y^* = [y_1^*, ..., y_m^*]$ is the corresponding reference summary of m tokens.

Policy Learning Several RL-based summarization (Kryściński et al., 2018; Li et al., 2018; Pasunuru and Bansal, 2018; Paulus et al., 2017) apply the self-critical policy gradient training algorithm (Rennie et al., 2017). Following (Paulus et al., 2017) we use REINFORCE algorithm, using as the baseline a greedy decoding algorithm according to the conditional distribution $p(y|X)$, giving rise to a sequence $\widehat{Y}$. The model is sampled using its Markov property, that is, one token at a time, giving rise to the sequence Y^s.

Following the standard RL actor-critic scheme, with $r(Y)$ the reward function for an output sequence Y, the loss to be *minimized* is then defined as:

$$L_{rl} = (r(\widehat{Y}) - r(Y^s)) \sum_{t=0}^{m} \log p(y_t^s | y_0^s, ..., y_{t-1}^s, X) \tag{4}$$

As ROUGE is the most widely used evaluation metric, Paulus et al. (2017) used ROUGE-L as the reward r for the L_{rl} function and tested the following three different setups:

- *ML*: the model trained with L_{ml} ($\gamma = 0$);
- *RL*: the model trained with L_{rl} ($\gamma = 1$);
- *ML+RL*: the model trained with L_{ml+rl} ($\gamma = 0.9984$).

The human evaluation conducted on the three models shows that *RL* performs worse than *ML*, and *ML+RL* performs best for both *readability* and *relevance*. The authors also conclude that "despite their common use for evaluation, ROUGE scores have their shortcomings and should not be the only metric to optimize on summarization model for long sequences", which is translated in the very high optimal γ. We show that using a more sensible metric to optimize leads to a better model, and to a lower γ.

4 Experiments

In our experiments, we evaluate the effect of substituting the ROUGE reward in the reinforcement-learning model of (Paulus et al., 2017) by our proposed metric (section 2). We, moreover, study the effect of using metrics that do not necessitate human-generated summaries.

4.1 Data Used

Task-specific corpora for building and evaluating summarization models associate a human-generated reference summary with each text provided. We resort to the CNN/Daily Mail (CNN-DM) dataset (Hermann et al., 2015; Nallapati et al., 2016) for our experiments. It includes 287,113 article/summary pairs for training, 13,368 for validation, and 11,490 for testing. The summary corresponding to each article consists of several bullet points displayed on the respective news outlet webpage. In average, summaries contain 66 tokens ($\sigma = 26$) and 4.9 bullet points. Consistently with See et al. (2017) and Gehrmann et al. (2018), we use the non-anonymized version of the dataset, the same training/validation splits, and perform truncation of source documents and summaries to 400 and 100 tokens, respectively.

To assess the possible benefits of reinforcing over the proposed QG-based metric, which does not require human-generated reference summaries, we employ TL;DR[2], a large-scale dataset for automatic summarization built on social media data, compounding to 4 Million training pairs (Völske et al., 2017). Both CNN-DM and TL;DR datasets are in English.

4.2 Models

For all our experiments, we build on top of the publicly available OpenNMT implementation[3], consistently with Gehrmann et al. (2018) to which we refer to as a baseline. The encoder is composed of a one-layer bi-LSTM with 512 hidden states, and 512 hidden states for the one-layer decoder. The embedding size is set at 128. The model is trained with Adagrad, with an initial learning rate of 0.15, and an initial accumulator value of 0.1. We continue training until convergence; when the validation perplexity does not decrease after an epoch, the learning rate is halved. We use gradient-clipping with a maximum norm of 2.

Gehrmann et al. (2018) showed that increasing the number of hidden states leads to slight improvements in performance, at the cost of increased training time; thus, as reinforcement learning is computationally expensive, we build on top of the smallest model – nonetheless, we include the largest model by Gehrmann et al. (2018) in our discussion of results.

All the experimented reinforcement approaches use the mixed training objectives defined in equation 2, with the ML part corresponding to the previously described baseline model pretrained on the CNN-DM dataset. Compared models differ on the considered reward signals. They also differ on their use of additional unsupervised data, either *In-Domain* or *Out-of-Domain*, as discussed below.

4.2.1 Reward Signals

The three reward signals used throughout our experiments, are detailed below:

1. **ROUGE**: We use only ROUGE-L as reward signal within the baseline architecture, consistently with Paulus et al. (2017);

2. **QA**$_{learned}$: Conversely, we compute the reward by applying the learned coefficients to the three components of the learned metric, as obtained in Section 2.4.

3. **QA**$_{equally}$: We apply the mixed training objective function, using as a reward the three metric components of the learned metric (ROUGE-L, QA$_{conf}$, and QA$_{fscore}$) equally weighted: this corresponds to setting a value of 1 for α, β and δ in Eq. 1. This allows to see to which extent learning is sensitive to fitting human assessments.

For (2) and (3), we set γ (Eq. 2) to 0.5[4]. This shows that, compared to (Paulus et al., 2017), we do not need to use NLL to avoid the model from generating unreadable summaries.

4.2.2 In-Domain vs Out-of-Domain

Finally, we experiment with the proposed QA$_{conf}$ and QA$_{fscore}$ metrics in an unsupervised fashion, as they can be computed at article level – *i.e.* without accessing the reference human-generated summaries. We investigate the potential benefits of using this approach both *in-domain* and *out-of-domain*: for the former, we resort to the test set of the CNN-Daily Mail (CNN-DM) dataset; for the latter, we leverage the TL;DR corpus.

As the CNN-Daily Mail is built from mainstream news articles, and the TL;DR data comes from social media sources, we consider the latter

[2] https://tldr.webis.de

[3] http://opennmt.net/OpenNMT-py/Summarization.html

[4] We have run experiments with $\gamma = 0.5$, and $\gamma = 0.9984$ as Paulus et al. (2017); we report here the best performance which was obtained with the former.

as out-of-domain in comparison. From the latter, which includes circa 4 million samples, we randomly draw sample subsets of size comparable with CNN-DM for training, validation and testing splits.

Due to computational costs, we restrict these experiments to the model trained under reinforcement using the $\text{QA}_{learned}$ metric. Under this setup, the model has access at training time both to:

- *supervised* samples for which a reference summary is given (and thus all metrics, including ROUGE and NLL, can be computed as a training objective), coming from the training set of CNN-Daily Mail corpus ;
- *unsupervised* samples, for which no reference is available, thus allowing to only compute QA_{conf}(unsup) and QA_{fscore}(unsup). Three unsupervised settings are considered in the following:

  TL;DR, corresponding to the *out-of-domain* setting where we use articles from the *TL;DR* dataset;

  CNN-DM (VAL), corresponding to an *in-domain* setting where we use texts from the validation set from the CNN/Daily Mail dataset;

  and, *CNN-DM (TEST)* for an *in-domain* setting where we use the articles from the test set (thus containing texts used for evaluation purposes).

While all the data is from the CNN-DM train dataset in the *supervised* setups, for the *unsupervised* setups, we set the proportion of unsupervised data to 50% (either CNN-DM VAL, CNN-DM TEST for *in-domain* or TL;DR for *out-of-domain* data). Thus, for 50% of the data, the model has access only to the QA_{conf} and QA_{fscore} reward signals, since the ROUGE-L reward can only be computed on *supervised* batches.

Therefore, for all the unsupervised setups, in order to keep consistency in the reward signal throughout the training, we multiply by a factor of 2 the weight associated with ROUGE-L when this reward is computable, and set it to 0 otherwise.

4.3 Results

In Table 2, we report the results obtained from our experiments in comparison with previously proposed approaches. We observe that, unsurprisingly, reinforcing on ROUGE-L allows to obtain significant improvements over the state-of-the-art, in terms of ROUGE but at the cost of lower QA-based metrics. Conversely, reinforcing on the proposed metric improves consistently all its components (ROUGE-L, QA_{conf} and QA_{fscore}).

However, increasing the reward does not necessarily correlate with better summaries. The human inspection as reported by (Paulus et al., 2017) shows that the generated summaries reinforced on ROUGE-L are consistently on the low end in terms of readability and relevance.

A closer inspection of the generated summaries revealed that the sequences generated by this model seem to qualitatively degrade as the number of produced tokens grows: they often start with a reasonable sub-sequence, but quickly diverge towards meaningless outputs. This can be explained by the aforementioned drawbacks of ROUGE, which are likely amplified when used both as evaluation and reward: the system might be optimizing for ROUGE, at the price of losing the information captured with the NLL loss by its language model.

We hence conducted a human evaluation for the different setups, reported in Table 3, assessing their outputs for *readability* and *relevance* in line with Paulus et al. (2017). We randomly sampled 50 articles from the CNN-DM test set; since the learned metric used in our experiments is derived from the subset manually evaluated in Paulus et al. (2017) we ensured that there was no overlap with it. For each of those 50 articles, three English speakers evaluated the summaries generated by the 7 different systems reported in Table 2.

We observe that reinforcing using the proposed metric – which includes QA based metrics, leads to comparable performance in terms of ROUGE w.r.t. state-of-the-art approaches, while clear benefits emerge from the results of the human evaluation: a significant improvement in terms of relevance, particularly when leveraging *in-domain* data in an *unsupervised* setup. Not surprisingly, we observe an improvement for our model when reinforced through the learned metric compared to the one equally weighted. The slightly lower relevance scores observed for the $\text{QA}_{learned}$ w.r.t. $\text{QA}_{equally}$ are consistent with the lower ROUGE-L and QA_{fscore} reported in Table 2. This is explained by the lower coefficients for ROUGE-L

	R-1	R-2	R-L	QA$_{fscore}$	QA$_{conf}$
See et al. (2017)	39.53	17.28	36.38	-	-
Gehrmann et al. (2018)	41.22	**18.68**	38.34	-	-
ML+RL Paulus et al. (2017)	39.87	15.82	36.90	-	-
RL Paulus et al. (2017)	41.16	15.75	39.08	-	-
Pasunuru and Bansal (2018)	40.43	18.00	37.10	-	-
Chen and Bansal (2018)	40.88	17.80	38.54	-	-
baseline	42.24	17.78	37.44	14.91	40.12
+ ROUGE	**45.62**	16.30	**41.60**	13.64	37.90
+ QA$_{equally}$	43.36	18.06	38.33	16.06	41.01
+ QA$_{learned}$	42.71	17.81	37.94	15.19	41.39
+ QA$_{learned}$ + TL;DR	42.75	17.57	37.88	15.75	41.54
+ QA$_{learned}$ + CNN-DM (VAL)	43.00	17.66	38.23	16.16	41.75
+ QA$_{learned}$ + CNN-DM (TEST)	42.74	17.25	37.96	**16.17**	**42.14**

Table 2: Comparison with previous works. On top, we report the results obtained by Gehrmann et al. (2018) using their largest architecture, as well as those by See et al. (2017). Next, we report results recently obtained by reinforcement learning approaches. Finally, we indicate the scores obtained by our baseline – the "small" model by Gehrmann et al. (2018) – and the six reinforced models we build on top of it.

and QA$_{fscore}$ (see 2.4), and the relatively stronger correlation of those two metrics with *relevance* (see Table 1).

Consistently with the figures reported in Table 2, the human evaluation results – reported in Tables 3 and 4 – confirm the progressive improvements of our different proposed models when using unsupervised data closer to the test set documents:

- adding *unsupervised* data from the out-of-domain TL;DR brings a slight improvement using QA$_{learned}$;
- when it comes to the same domain (*i.e.* CNN-DM validation) the improvements increase;
- finally, when unsupervised samples come from the same set as those used for testing, we observe even better results.

These results show that using the proposed QA-based metrics, that do not depend on reference summaries, allows to leverage raw text data; and, that fine-tuning (without supervision) on the documents to be summarized is beneficial.

To elaborate further, we notice that applying the learned coefficients for 1 to the results obtained by models reinforced on QA$_{\text{learned}}$ and QA$_{\text{equally}}$, see Table 2, we obtain very similar scores (namely, 136.43 for QA$_{\text{equally}}$ and 136.4 for QA$_{\text{learned}}$). However, the qualitative analysis reported in Tables 3 and 4 shows that while they perform similarly in terms of *relevance*, a significantly lower score for *readability* is obtained using QA$_{\text{equally}}$. This can be explained by the stronger weight of ROUGE_L for this setup, a fact which might lead to a degradation of the quality of the output consistently with the observations reported in (Paulus et al., 2017) as well as in our ROUGE experiment.

Another observation from Tables 3 and 4 is that while QA$_{\text{learned}}$ performs significantly better in term of *readability* than QA$_{\text{learned}}$ + CNN-DM (VAL), the opposite holds for *relevance*. This could be explained by the setup difference during training: as detailed in section 4.2.2, for unsupervised setups (*i.e.* QA$_{\text{learned}}$ + CNN-DM (VAL)) only the QA-based metrics are computed for the portion of data for which no reference is available. While testing (TEST) and validation (VAL) splits come the same dataset (CNN-DM), we observe that using the samples from TEST in an unsupervised fashion allows for maintaining comparably high *relevance* compared to QA$_{\text{learned}}$ + CNN-DM (VAL), while also obtaining similar *readability* to QA$_{\text{learned}}$. This shows the possible benefits that can be obtained by exposing the model to the evaluation data in unsupervised setups. To further study our unsupervised metrics, we performed additional experiments on the TL;DR corpus. We observed more than one absolute point of improvement w.r.t CNN-DM TEST in terms of ROUGE-L, QA$_{fscore}$ (unsup) and QA$_{conf}$ (unsup).

This indicates that the proposed unsupervised

	Readability	Relevance
human reference	7.27*	7.4**
baseline	7.07	5.82
+ ROUGE	2.14**	5.48**
+ QA$_{equally}$	5.94**	6.34**
+ QA$_{learned}$	6.96	6.21**
+ QA$_{learned}$ + TL;DR	6.60*	6.26**
+ QA$_{learned}$ + CNN-DM (VAL)	6.40*	6.75**
+ QA$_{learned}$ + CNN-DM (TEST)	6.89	6.80**

Table 3: Human assessment: two-tailed t-test results are reported for each model compared to the baseline ($* : p < .01$, $** : p < .001$).

	human reference	baseline	+ ROUGE	+ QA$_{equally}$	+ QA$_{learned}$	+ QA$_{learned}$ + TL;DR	+ QA$_{learned}$ + CNN-DM (VAL)	+ QA$_{learned}$ + CNN-DM (TEST)
baseline	* / **	-						
+ ROUGE	** / **	** / **	-					
+ QA$_{equally}$	** / **	** / **	** / **	-				
+ QA$_{learned}$	** / **	- / **	** / **	** / -	-			
+ QA$_{learned}$ + TL;DR	** / **	* / **	** / **	** / -	* / -	-		
+ QA$_{learned}$ + CNN-DM (VAL)	** / **	* / **	** / **	** / *	** / **	- / *	-	
+ QA$_{learned}$ + CNN-DM (TEST)	** / **	- / **	** / **	** / *	- / **	- / *	* / -	-

Table 4: Human assessment: two-tailed t-test results are reported for each model pair for Readability / Relevance ($* : p < .01$, $** : p < .001$).

metrics allow the model to better transfer to new domains such as TL;DR. These results pave the way for leveraging large numbers of texts, in a self-supervised manner, to train automatic summarization models.

5 Conclusions

We have presented the analysis of novel QA-based metrics[5], and have shown promising results when using them as a reward in a RL setup. Crucially, those metrics do not require a human reference, as they can be computed from the text to be summarized.

From our experiments this proves particularly beneficial, allowing to leverage both *in-domain* and *out-of-domain* unlabeled data.

The promising results obtained indicate a path towards partially self-supervised training of summarization models, and suggest that progress in automated question generation can bring benefits for automatic summarization.

Finally, to our knowledge, this paper is the first to compare two architectures with the same reinforcement setup on the same data: the one proposed by See et al. (2017) and extended by Gehrmann et al. (2018), versus the one by Paulus et al. (2017). In terms of ROUGE, we observe better results than those reported by Paulus et al. (2017) – see Table 2 – indicating a possible edge for the architecture proposed by See et al. (2017).

[5] A python package will be made available at https://www.github.com/recitalAI/summa-qa.

References

Kristjan Arumae and Fei Liu. 2019. Guiding extractive summarization with question-answering rewards. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2566–2577, Minneapolis, Minnesota. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ping Chen, Fei Wu, Tong Wang, and Wei Ding. 2018. A semantic qa-based approach for text summarization evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. Association for Computational Linguistics.

Piji Li, Lidong Bing, and Wai Lam. 2018. Actor-critic based training framework for abstractive summarization. *arXiv preprint arXiv:1803.11070*.

C-Y Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. of Workshop on Text Summarization Branches Out, Post Conference Workshop of ACL 2004*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Ani Nenkova. 2011. *Automatic Summarization*. Foundations and Trends in Information Retrieval. Now.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

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

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, Volume 2 (Short Papers)*, pages 646–653, New Orleans, Louisiana. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

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

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

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2018. Summarization evaluation in the absence of human model summaries using the compositionality of word embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 905–914, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl;dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Harris Wu, Dragomir R Radev, and Weiguo Fan. 2002. Towards answer-focused summarization. In *Proceedings of the 1st International Conference on Information Technology and Applications*.