Deep Reinforcement Learning with Distributional Semantic Rewards for Abstractive Summarization

Siyao Li\textsuperscript{1*}, Deren Lei\textsuperscript{1*}, Pengda Qin\textsuperscript{2}, William Yang Wang\textsuperscript{1}

\textsuperscript{1}University of California, Santa Barbara
\textsuperscript{2}Beijing University of Posts and Telecommunications
\{siyaoli, derenlei\}@ucsb.edu, qinpengda@bupt.edu.cn, william@cs.ucsb.edu

Abstract

Deep reinforcement learning (RL) has been a commonly-used strategy for the abstractive summarization task to address both the exposure bias and non-differentiable task issues. However, the conventional reward \textsc{ROUGE-L} simply looks for exact n-grams matches between candidates and annotated references, which inevitably makes the generated sentences repetitive and incoherent. In this paper, instead of \textsc{ROUGE-L}, we explore the practicability of utilizing the distributional semantics to measure the matching degrees. With distributional semantics, sentence-level evaluation can be obtained, and semantically-correct phrases can also be generated without being limited to the surface form of the reference sentences. Human judgments on Gigaword and CNN/Daily Mail datasets show that our proposed distributional semantics reward (DSR) has distinct superiority in capturing the lexical and compositional diversity of natural language.

1 Introduction

Abstractive summarization is a task of paraphrasing a long article with fewer words. Unlike extractive summarization, abstractive summaries can include tokens out of the article’s vocabulary. There exists several encoder-decoder approaches such as attention-based architecture (Bahdanau et al., 2015; Rush et al., 2015; Nallapati et al., 2016; Chopra et al., 2016), incorporating graph techniques (Moawad and Aref, 2012; Ganesan et al., 2010), and adding pointer generator (Vinyals et al., 2015; Nallapati et al., 2016; See et al., 2017; Bello et al., 2017). However, long sentence generation suffers from exposure bias (Bahdanau et al., 2017) as the error accumulates during the decoding process.

\*Equal contributions.

Many innovative deep RL methods (Ranzato et al., 2016; Wu et al., 2016; Paulus et al., 2018; Lamb et al., 2016) are developed to alleviate this issue by providing sentence-level feedback after generating a complete sentence, in addition to optimal transport usage (Napoles et al., 2012). However, commonly used automatic evaluation metrics for generating sentence-level rewards count exact n-grams matches and are not robust to different words that share similar meanings since the semantic level reward is deficient.

Currently, many studies on contextualized word representations (Peters et al., 2018; Devlin et al., 2019) prove that they have a powerful capacity of reflecting distributional semantic. In this paper, we propose to use the distributional semantic reward to boost the RL-based abstractive summarization system. Moreover, we design several novel objective functions.

Experiment results show that they outperform the conventional objectives while increasing the sentence fluency. Our main contributions are three-fold:

\begin{itemize}
  \item We are the first to introduce DSR to abstractive summarization and achieve better results than conventional rewards.
  \item Unlike \textsc{ROUGE}, our DSR does not rely on cross-entropy loss (XENT) to produce readable phrases. Thus, no exposure bias is introduced.
  \item DSR improves generated tokens’ diversity and fluency while avoiding unnecessary repetitions.
\end{itemize}

2 Methodology

Background While sequence models are usually trained using XENT, they are typically evaluated at test time using discrete NLP metrics such as \textsc{BLEU} (Papineni et al., 2002), \textsc{ROUGE} (Lin, 2004), \textsc{METEOR} (Banerjee and Lavie, 2005). Therefore, they suffer from both the exposure bias...
and non-differentiable task metric issues. To solve these problems, many resorts to deep RL with sequence-to-sequence model (Paulus et al., 2018; Ranzato et al., 2016; Ryang and Abekawa, 2012), where the learning agent interacts with a given environment. However, RL models have poor sample efficiency and lead to very slow convergence rate. Therefore, RL methods usually start from a pretrained policy, which is established by optimizing XENT at each word generation step.

\[ L_{XENT} = -\sum_{t=1}^{n'} \log P(y_t|y_1, \ldots, y_{t-1}, x) \]  

(1)

Then, during RL stage, the conventional way is to adopt self-critical strategy to fine-tune based on the target evaluation metric,

\[ L_{RL} = \sum_{t=1}^{n'} \log P(\hat{y}_t|\hat{y}_1, \ldots, \hat{y}_{t-1}, x) \times (r_{metric}(\hat{y}^b_t) - r_{metric}(\hat{y}_t)) \]  

(2)

### Distributional Semantic Reward

During evaluating the quality of the generated sentences, ROUGE looks for exact matches between references and generations, which naturally overlooks the expression diversity of the natural language. In other words, it fails to capture the semantic relation between similar words. To solve this problem, distributional semantic representations are a practical way. Recent works on contextualized word representations, including ELMO (Peters et al., 2018), GPT (Radford et al., 2018), BERT (Devlin et al., 2019), prove that distributional semantics can be captured effectively. Based on that, a recent study, called BERTSCORE (Zhang et al., 2019), focuses on sentence-level generation evaluation by using pre-trained BERT contextualized embeddings to compute the similarity between two sentences as a weighted aggregation of cosine similarities between their tokens. It has a higher correlation with human evaluation on text generation tasks comparing to existing evaluation metrics.

In this paper, we introduce it as a DSR for deep RL. The BERTSCORE is defined as:

\[ R_{BERT} = \frac{\sum_{y_j \in y} \text{idf}(y_j) \max_{y_i \in y} y_i^\top \hat{y}_j}{\sum_{y_i \in y} \text{idf}(y_i)} \]  

(3)

\[ P_{BERT} = \frac{\sum_{y_j \in y} \text{idf}(y_j) \max_{y_i \in y} y_j^\top \hat{y}_i}{\sum_{y_j \in y} \text{idf}(y_j)} \]  

(4)

\[ F_{BERT} = \frac{2R_{BERT} \cdot P_{BERT}}{R_{BERT} + P_{BERT}} \]  

(5)

where \( y \) and \( \hat{y} \) represent BERT contextual embeddings of reference word \( y \) and candidate word \( \hat{y} \), respectively. The function \( \text{idf}(\cdot) \) calculates inverse document frequency (idf). In our DSR, we do not use the idf since Zhang et al. (2019) requires to use the entire dataset including test set for calculation. Besides, ROUGE do not use similar weight, so we do not include idf for consistency.

### 3 Experimental Setup

#### 3.1 Datasets

**Gigaword corpus** It is an English sentence summarization dataset based on annotated Gigaword (Napoles et al., 2012). A single sentence summarization is paired with a short article. We use the OpenNMT provided version It contains 3.8M training, 189k development instances. We randomly sample 15k instances as our test data.

**CNN/Daily Mail dataset** It consists of online news articles and their corresponding multi-sentence abstracts (Hermann et al., 2015; Nallapati et al., 2016). We use the non-anonymized version provided by See et al. (2017), which contains 287k training, 13k validation, and 11k testing examples. We truncate the articles to 400 tokens and limit the summary lengths to 100 tokens.

#### 3.2 Pretrain

We first pretrain a sequence-to-sequence model with attention using XENT and then select the best parameters to initialize models for RL. Our models have 256-dimensional hidden states and 128-dimensional word embeddings and also incorporates the pointer mechanism (See et al., 2017) for handling out of vocabulary words.

#### 3.3 Baseline

In abstractive summarization, ROUGE (Lin, 2004) is a common evaluation metric to provide a sentence-level reward for RL. However, using ROUGE as a pure RL objective may cause too many repetitions and reduced fluency in outputs. Paulus et al. (2018) propose a hybrid learning objective that combines XENT and self-critical ROUGE reward (Paulus et al., 2018).

\[ L_{baseline} = \gamma L_{rouge} + (1 - \gamma)L_{XENT} \]  

(6)
where $\gamma$ is a scaling factor and the F score of ROUGE-L is used as the reward to calculate $L_{\text{Rouge}}$. In our experiment, we select $\gamma = 0.998$ for Gigaword Corpus and $\gamma = 0.9984$ for CNN/Daily Mail dataset. Note that we do not avoid repetition during the test time as Paulus et al. (2018) do, because we want to examine the repetition of sentence directly produced after training.

### 3.4 Proposed Objective Functions

Inspired by the above objective function (Paulus et al., 2018), we optimize RL models with a similar loss function as equation 2. Instead of ROUGE-L, we incorporate BERTSCORE, a DSR to provide sentence-level feedback.

In our experiment, $L_{\text{DSR}}$ is the self-critical RL loss (equation 2) with $F_{\text{BERT}}$ as the reward. We introduce the following objective functions:

**DSR+ROUGE:** A combined reward function of ROUGE and $F_{\text{BERT}}$

$$L_1 = \gamma L_{\text{DSR}} + (1 - \gamma)L_{\text{Rouge}}$$  \hspace{1cm} (7)

In our experiment, we select $\gamma = 0.5$ for both datasets to balance the influence of two reward functions.

**DSR+XENT:** $F_{\text{BERT}}$ reward with XENT to make the generated phrases more readable.

$$L_2 = \gamma' L_{\text{DSR}} + (1 - \gamma')L_{\text{XENT}}$$  \hspace{1cm} (8)

In our experiment, we select $\gamma' = 0.998$ for Gigaword Corpus and $\gamma' = 0.9984$ for CNN/Daily Mail dataset.

**DSR:** Pure $F_{\text{BERT}}$ objective function without any teacher forcing.

$$L_3 = L_{\text{DSR}}$$  \hspace{1cm} (9)

### 4 Results

For the abstractive summarization task, we test our models with different objectives on the Gigaword and CNN/Daily Mail datasets. We choose $F_{\text{BERT}}$ and ROUGE-L as our automatic evaluation metrics. For multi-sentence summaries in CNN/Daily Mail dataset, we calculate ROUGE-L and $F_{\text{BERT}}$ after concatenating the original abstract sentences or their BERT contextualized embeddings. The results are shown in Table 1. The model DSR+ROUGE (equation 7) achieves the best ROUGE performance on Gigaword and similar performance comparing to ROUGE model on CNN/Daily Mail. Training with DSR (equation 9) and DSR+XENT (equation 8) can obtain the best BERTScores as expected. It is also expected that ROUGE model will obtain worse BERTScores as simply optimizing ROUGE will generate less readable sentences (Paulus et al., 2018); however, DSR model without XENT as a teacher forcing can improve the performance of pretrained model in both $F_{\text{BERT}}$ and ROUGE-L scale.

Note that DSR model’s ROUGE-L is high in training time but does not have a good generalization on test set, and ROUGE-L is not our target evaluation metrics. In the next section, we will do human evaluation to analyze the summarization performance of different reward systems.

### 5 Human Evaluation

We perform human evaluation on the Amazon Mechanical Turk to assure the benefits of DSR on output sentences’ coherence and fluency. We randomly sample 500 items as an evaluation set us-
Table 1: Results on Gigawords and CNN/Daily Mail for abstractive summarization. Upper rows show the results of baselines. Rouge stands for the F score of Rouge-L.

<table>
<thead>
<tr>
<th>Model</th>
<th>Gigawords</th>
<th>CNN/Daily Mail</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F_{BERT}$</td>
<td>$P_{BERT}$</td>
</tr>
<tr>
<td>XENT</td>
<td>65.78</td>
<td>67.61</td>
</tr>
<tr>
<td>ROUGE</td>
<td>61.46</td>
<td>60.89</td>
</tr>
<tr>
<td>ROUGE+XENT</td>
<td>66.50</td>
<td>67.24</td>
</tr>
<tr>
<td>DSR+ROUGE</td>
<td>66.48</td>
<td>66.79</td>
</tr>
<tr>
<td>DSR+XENT</td>
<td>67.02</td>
<td>67.69</td>
</tr>
<tr>
<td>DSR</td>
<td>67.06</td>
<td>67.34</td>
</tr>
</tbody>
</table>

Table 2: Human evaluation results on Gigaword and CNN/Daily Mail for abstractive summarization. ROUGE+XENT is the selected baseline model.

<table>
<thead>
<tr>
<th>Task</th>
<th>Model v. base</th>
<th>Gigawords</th>
<th>CNN/Daily Mail</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Win</td>
<td>Lose</td>
</tr>
<tr>
<td>Relevance</td>
<td>DSR + XENT</td>
<td>31.0%</td>
<td>25.2%</td>
</tr>
<tr>
<td></td>
<td>DSR</td>
<td>45.4%</td>
<td>27.2%</td>
</tr>
<tr>
<td>Fluency</td>
<td>DSR + XENT</td>
<td>40.2%</td>
<td>19.6%</td>
</tr>
<tr>
<td></td>
<td>DSR</td>
<td>45.4%</td>
<td>28.8%</td>
</tr>
</tbody>
</table>

Table 3: Qualitative analysis on repetition(Rep) / diversity(Div). They are calculated by the percentage of repeat/out-of-article n-grams (unigrams for Gigaword and 5-grams for CNN/Daily Mail) in generated sentences.

<table>
<thead>
<tr>
<th>Model</th>
<th>Gigawords</th>
<th>CNN/Daily Mail</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rep(%)</td>
<td>Div(%)</td>
</tr>
<tr>
<td>ROUGE+XENT</td>
<td>11.57</td>
<td>16.33</td>
</tr>
<tr>
<td>DSR+ROUGE</td>
<td>18.42</td>
<td>15.97</td>
</tr>
<tr>
<td>DSR+XENT</td>
<td>20.15</td>
<td>16.24</td>
</tr>
<tr>
<td>DSR</td>
<td>7.20</td>
<td>19.17</td>
</tr>
</tbody>
</table>

6 Analysis

Diversity Other than extractive summarization, abstractive summarization allows more degrees of freedom in the choice of words. While simply selecting words from the article made the task easier to train, higher action space can provide more paths to potentially better results (Nema et al., 2017). Using the DSR as deep RL reward will support models to choose actions that are not n-grams of the articles. In Table 4, we list a few generated samples on Gigaword Corpus. In our first example in Table 4, the word “sino-german” provides an interesting and efficient way to express the relation between China and Germany. $F_{BERT}$ is also improved by making this change. In addition, the second example in Table 4 shows that RL model with DSR corrects the sentence’ grammar and significantly improves the $F_{BERT}$ score by switching “down” to an unseen word “drops”. On the other hand, when optimizing DSR to improve the diversity of generation, some semantically similar words may also be generated and harm the summarization quality as shown in the third example in Table 4. The new token “wins” reduces the scores of both metrics. We also evaluate the diversity of a model quantitively by averaging the percentage of out-of-article n-grams in generated sentences. Results can be found in Table 3. The DSR model achieves the highest diversity.
### Table 4: Qualitative analysis of generated samples on Gigaword corpus. Generated words that do not appear in the context are marked blue. Repeated words are marked red. The first two examples represent DSR’s generated tokens are more diverse. However, it may suffer from problems as shown in example 3 and 4.

#### Repetition
Repetition will lead to lower $F_B$ERT as shown in the last example in Table 4. Using DSR reduces the probability of producing repetitions. The average percentage of repeated n-grams in generated sentences are presented in the Table 3. As shown in this table, unlike ROUGE, the DSR model can achieve high fluency without XENT; moreover, it produces the fewest repetitions among all the rewards. Table 4 gives an example that DSR produces a repeated word (from example 4), but it does not reflect the overall distribution of repeated word generation for all evaluated models.

#### 7 Conclusion
This paper demonstrates the effectiveness of applying the distributional semantic reward to reinforcement learning in abstractive summarization, and specifically, we choose BERTSCORE. Our experimental results demonstrate that we achieve better performance on Gigaword and CNN/Daily Mail datasets. Besides, the generated sentences have fewer repetitions, and the fluency is also improved. Our finding is aligned to a contemporaneous study (Wieting et al., 2019) on leveraging semantic similarity for machine translation.


