Submodular Optimization-based Diverse Paraphrasing and its Effectiveness in Data Augmentation

Ashutosh Kumar[*1] **Satwik Bhattamishra**[*2] [†] **Manik Bhandari**[1] **Partha Talukdar**[1]
[1] Indian Institute of Science, Bangalore
[2] Birla Institute of Technology and Science, Pilani
ashutosh@iisc.ac.in, satwik55@gmail.com, mbbhandarimanik@gmail.com, ppt@iisc.ac.in

Abstract

Inducing diversity in the task of paraphrasing is an important problem in NLP with applications in data augmentation and conversational agents. Previous paraphrasing approaches have mainly focused on the issue of generating semantically similar paraphrases, while paying little attention towards diversity. In fact, most of the methods rely solely on top-k beam search sequences to obtain a set of paraphrases. The resulting set, however, contains many structurally similar sentences. In this work, we focus on the task of obtaining highly diverse paraphrases while not compromising on paraphrasing quality. We provide a novel formulation of the problem in terms of monotone submodular function maximization, specifically targeted towards the task of paraphrasing. Additionally, we demonstrate the effectiveness of our method for data augmentation on multiple tasks such as intent classification and paraphrase recognition. In order to drive further research, we have made the source code available.

1 Introduction

Paraphrasing is the task of rephrasing a given text in multiple ways such that the semantics of the generated sentences remain unaltered. Paraphrasing *Quality* can be attributed to two key characteristics - *fidelity* which measures the semantic similarity between the input text and generated text, and *diversity*, which measures the lexical dissimilarity between generated sentences.

Many previous works (Prakash et al., 2016; Gupta et al., 2018; Li et al., 2018) address the task of obtaining semantically similar paraphrases. While it is essential to produce paraphrases with high fidelity, it is equally important, and in many cases desirable, to produce lexically diverse ones. Diversity in paraphrase generation finds applications in text simplification (Nisioi et al., 2017; Xu et al., 2015), document summarization (Li et al., 2009; Nema et al., 2017), QA systems (Fader et al., 2013; Bernhard and Gurevych, 2008), data augmentation (Zhang et al., 2015; Wang and Yang, 2015), conversational agents (Li et al., 2016) and information retrieval (Anick and Tipirneni, 1999).

SOURCE	– how do i increase body height ?
REFERENCE	– what do i do to increase my height ?
BEAM SEARCH	– how do i increase my height ? – how do i increase my body height ? – how do i increase the height ? – how would i increase my body height ?
DIPS (OURS)	– how could i increase my height ? – what should i do to increase my height ? – what are the fastest ways to increase my height ? – is there any proven method to increase height ?

Table 1: Sample paraphrases generated by beam search and DiPS (our method). It can be seen that DiPS offers lexically diverse paraphrases without compromising on fidelity.

To obtain a set of multiple paraphrases, most of the current paraphrasing models rely solely on top-k beam search sequences. The resulting set, however, contains many structurally similar sentences with only minor, word level changes. There have been some prior works (Li and Jurafsky, 2016; Elhamifar et al., 2012) which address the notion of diversity in NLP, including in sequence learning frameworks (Song et al., 2018; Vijayakumar et al., 2018). Although Song et al. (2018) address the issue of diversity in the scenario of neural conversation models using determinantal point processes (DPP), it could be naturally used for paraphrasing. On similar lines, subset selection based on Simultaneous Sparse Recovery (SSR) (Elhamifar et al., 2012) can also be easily adapted for the same task.

Though these methods are helpful in maximizing diversity, they are restrictive in terms of re-

[*]Equal Contribution
[†]This research was conducted during the author's internship at the Indian Institute of Science, Bangalore.

taining fidelity with respect to the source sentence. Addressing the task of diverse paraphrasing through the lens of monotone submodular function maximization (Fujishige, 2005; Krause and Golovin; Bach et al., 2013) alleviates this problem and also provides a few additional benefits. Firstly, the submodular objective offers better flexibility in terms of controlling diversity as well as fidelity. Secondly, there exists a simple greedy algorithm for solving monotone submodular function maximization (Nemhauser et al., 1978), which guarantees the diverse solution to be almost as good as the optimal solution. Finally, many submodular programs are fast and scalable to large datasets.

Below, we list the main contributions of our paper.

1. We introduce **Di**verse **P**araphraser using **S**ubmodularity (DiPS). DiPS maximizes a novel submodular objective function specifically targeted towards paraphrasing.
2. We perform extensive experiments to show the effectiveness of our method in generating structurally diverse paraphrases without compromising on fidelity. We also compare against several possible diversity inducing schemes.
3. We demonstrate the utility of diverse paraphrases generated via DiPS as data augmentation schemes on multiple tasks such as intent and question classification.

We have made DiPS's source code available at https://github.com/malllabiisc/DiPS

2 Related Work

Paraphrasing a given sentence is an important problem and numerous approaches have been proposed to address it. Recently *sequence-to-sequence* based data-driven deep learning models have been proposed, which try to address the limitations of earlier traditional rule-based (McKeown, 1983) methods. Prakash et al. (2016) employ residual connections in LSTM to enhance the traditional sequence-to-sequence model. Gupta et al. (2018) provide a VAE (Kingma and Welling, 2013) based framework to improve the quality of generated paraphrases. Li et al. (2018) propose a reinforcement learning based model which uses pointer-generator (See et al., 2017) for generating paraphrases and an evaluator based on (Parikh et al., 2016) to penalize non-paraphrastic generations. Several other works (Cao et al., 2017; Iyyer et al., 2018) exist for paraphrasing, though they have either been superseded by newer models or are not-directly applicable to our settings. However, most of these methods focus on the issue of generating semantically similar paraphrases, while paying little attention towards diversity.

Diversity in paraphrasing models was first explored by (Gupta et al., 2018) where they propose to generate variations based on different samples from the latent space in a deep generative framework. Although diversity in paraphrasing models has not been explored extensively, methods have been proposed to address diversity in other NLP tasks (Li et al., 2016, 2015; Gimpel et al., 2013). Diverse beam search proposed by (Vijayakumar et al., 2018) generates k-diverse sequences by dividing the candidate subsequences at each time step into several groups and penalizing subsequences which are similar to prior groups. The most relevant to our approach is the method proposed by (Song et al., 2018) for neural conversation models where they incorporate diversity by using DPP to select diverse subsequences at each time step. Although their work is addressed in the scenario of neural conversation models, it could be naturally adapted to paraphrasing models and thus we use it as a baseline.

Submodular functions have been applied to a wide variety of problems in machine learning (Iyer and Bilmes, 2013; Jegelka and Bilmes, 2011; Krause and Guestrin, 2011; Kolmogorov and Zabih, 2002) and have recently attracted much attention in several NLP tasks including document summarization (Lin and Bilmes, 2011), data selection in machine translation (Kirchhoff and Bilmes, 2014) and goal-oriented chatbot training (Dimovski et al., 2018). However, their application to sequence generation is largely unexplored.

Data augmentation is a technique for increasing the size of labeled training sets by leveraging task specific transformations which preserve class labels. While the technique is ubiquitous in the vision community (Krizhevsky et al., 2012; Ratner et al., 2017), data-augmentation in NLP is largely under-explored. Most current augmentation schemes involve thesaurus based synonym replacement (Zhang et al., 2015; Wang and Yang, 2015), and replacement by words with paradigmatic relations (Kobayashi, 2018). Both of these

Algorithm 1: Greedy selection for submodular optimization (Cardinality constraint)

Input: Ground Set: V
Budget: k
Submodular Function: $\mathcal{F}$

$S \leftarrow \emptyset$
$N \leftarrow V$
while $|S| < k$ **do**
  $x^* \leftarrow \text{argmax}_{x \in N} \mathcal{F}(S \cup \{x\})$
  $S \leftarrow S \cup \{x^*\}$
  $N \leftarrow N \setminus \{x^*\}$
end
return S

Algorithm 2: DiPS

Input: Input Sentence: S_{in}
Max. decoding length: T
Submodular objective: $\mathcal{F}$
No. of paraphrases required: k

Process S_{in} using the encoder of SEQ2SEQ
Start the decoder with input symbol `sos`
$t \leftarrow 0$; $P \leftarrow \emptyset$
while $t < T$ **do**
  Generate top $3k$ most probable subsequences
  $P \leftarrow$ Select k based on $\text{argmax}_{X \subseteq V^{(t)}} \mathcal{F}(X)$ using **Algorithm 1**
  $t = t + 1$
end
return P

approaches try to boost the generalization abilities of downstream classification models through word-level substitutions. However, they are inherently restrictive in terms of the diversity they can offer. Our work offers a data-augmentation scheme via high quality paraphrases.

3 Background: Submodularity

Let V = $\{v_1, \ldots, v_n\}$ be a set of objects, which we refer to as the ground set, and $\mathcal{F} : 2^V \rightarrow \mathbb{R}$ be a set function which works on subsets S of V to return a real value. The task is to find a subset S of bounded cardinality say $|S| \leq k$ that maximizes the function $\mathcal{F}$, i.e., $\text{argmax}_{S \subseteq V} \mathcal{F}(S)$. In general, solving this problem is intractable. However, if the function $\mathcal{F}$ is monotone non-decreasing submodular, then although the problem is still NP-complete, there exists a greedy algorithm (Algorithm 1) (Nemhauser et al., 1978) that finds an approximate solution which is guaranteed to be within 0.632 of the optimal solution.

Submodular functions are set functions $\mathcal{F} : 2^V \rightarrow \mathbb{R}$, where 2^V denotes the power set of ground set V. Submodular functions satisfy the following equivalent properties of *diminishing returns*: $\forall X, Y \subseteq V$ with $X \subseteq Y$, and $\forall s \in V \setminus Y$, we have the following.

$$\mathcal{F}(X \cup \{s\}) - \mathcal{F}(X) \geq \mathcal{F}(Y \cup \{s\}) - \mathcal{F}(Y) \quad (1)$$

In other words, the *value* addition due to incorporation of s decreases as the subset grows from X to Y. Equivalently, $\forall X, Y \subseteq V$, we have,

$$\mathcal{F}(X) + \mathcal{F}(Y) \geq \mathcal{F}(X \cup Y) + \mathcal{F}(X \cap Y)$$

In case the above inequalities are equalities, the function $\mathcal{F}$ is said to be modular. Let $\mathcal{F}(s|X) \triangleq \mathcal{F}(X \cup \{s\}) - \mathcal{F}(X)$. Therefore, $\mathcal{F}$ is submodular if $\mathcal{F}(s|X) \geq \mathcal{F}(s|Y)$ for $X \subseteq Y$.

The second criteria which the function needs to satisfy for Algorithm 1 to be applicable is of monotonicity. A set function $\mathcal{F}$ is said to be monotone non-decreasing if $\forall X \subseteq Y, \mathcal{F}(X) \leq \mathcal{F}(Y)$.

Submodular functions are relevant in a large class of real-world applications, therefore making them extremely useful in practice. Additionally, submodular functions share many commonalities with convex functions, in the sense that they are closed under a number of standard operations like mixtures (non-negative weighted sum of submodular functions), truncation and some restricted compositions.

The above properties will be useful when defining the submodular objective for obtaining high quality paraphrases.

4 Methodology

Similar to Prakash et al. (2016); Gupta et al. (2018); Li et al. (2018), we formulate the task of paraphrase generation as a sequence-to-sequence learning problem. Previous SEQ2SEQ based approaches depend entirely on the standard cross-entropy loss to produce semantically similar sentences and greedy decoding during generation. However, this does not guarantee lexical variety in the generated paraphrases. To incorporate some form of diversity, most prior approaches rely solely on top-k beam search sequences. The k-best list generated by standard beam search are a poor surrogate for the entire search space (Finkel et al., 2006). In fact, most of the sentences in the resulting set are structurally similar, differing only by punctuations or minor morphological variations.

While being similar in the encoding scheme, our work adopts a different approach for the final decoding. We propose a framework which organi-

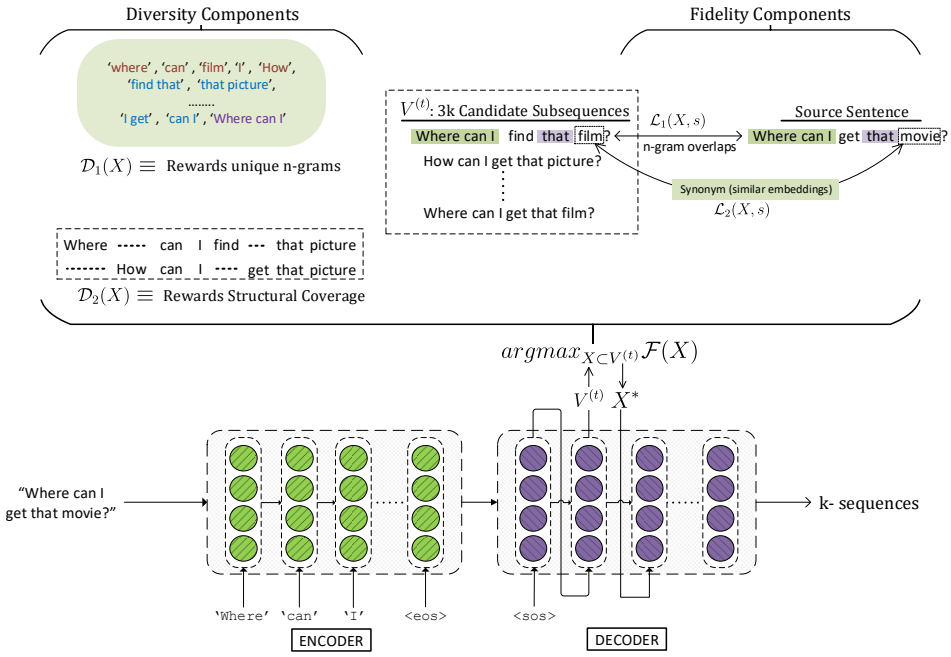

Figure 1: Overview of DiPS during decoding to generate k paraphrases. At each time step, a set of N sequences ($V^{(t)}$) is used to determine $k < N$ sequences (X^*) via submodular maximization . The above figure illustrates the motivation behind each submodular component. Please see Section 4 for details.

cally combines a sentence encoder with a diversity inducing decoder.

4.1 Overview

Our approach is built upon SEQ2SEQ framework. We first feed the tokenized source sentence to the encoder. The task of the decoder is to take as input the encoded representation and produce the respective paraphrase. To achieve this, we train the model using standard cross entropy loss between the generated sequence and the target paraphrase. Upon completion of training, instead of using greedy decoding or standard beam search, we use a modified decoder where we incorporate a submodular objective to obtain high quality paraphrases. Please refer to Figure 1 for an overview of the proposed method.

During the generation phase, the encoder takes the source sentence as input and feeds its representation to the decoder to initiate the decoding process. At each time-step t, we consider N most probable subsequences since they are likely to be well-formed. Based on optimization of our submodular objective, a subset of size $k < N$ are selected and sent as input to the next time step $t + 1$ for further generation. The process is repeated until desired output length T or <eos> token, whichever comes first.

4.2 Monotone Submodular Objectives

We design a parameterized class of submodular functions tailored towards the task of paraphrase generation. Let $V^{(t)}$ be the ground set of possible subsequences at time step t. We aim to determine a set $X \subseteq V^{(t)}$ that retains certain *fidelity* as well as *diversity*. Hence, we model our submodular objective function as follows:

$$X^* = \underset{X \subseteq V^{(t)}}{\operatorname{argmax}} \mathcal{F}(X) \quad s.t. \ |X| \leq k \qquad (2)$$

where k is our budget (desired number of paraphrases) and $\mathcal{F}$ is defined as:

$$\mathcal{F}(X) = \lambda \mathcal{L}(X, s) + (1 - \lambda)\mathcal{D}(X) \qquad (3)$$

Here s is the source sentence, $\mathcal{L}(X, s)$ and $\mathcal{D}(X)$ measure *fidelity* and *diversity*, respectively. $\lambda \in [0, 1]$ is the trade-off coefficient. This formulation clearly brings out the trade-off between the two key characteristics.

Fidelity

It is imperative to design functions that exploit the decoder search space to maximize the semantic similarity between the generated and the source sentence. To achieve this we build upon a known

class of monotone submodular functions (Stobbe and Krause, 2010)

$$f(X) = \sum_{i \in U} \mu_i \phi_i(m_i(X)) \tag{4}$$

where U is the set of features to be defined later, $\mu_i \geq 0$ is the feature weight, $m_i(X) = \sum_{x \in X} m_i(x)$ is non-negative modular function and ϕ_i is a non-negative non-decreasing concave function. Based on the analysis of concave functions in (Kirchhoff and Bilmes, 2014), we use the simple square root function as ϕ ($\phi(a) = \sqrt{a}$) in both of our fidelity objectives defined below.

We consider two complementary notions of sentence similarity namely syntactic and semantic. To capture syntactic information we define the following function:

$$\mathcal{L}_1(X, s) = \mu_1 \sqrt{\sum_{x \in X} \sum_{n=1}^{N} \beta^n \left| x_{\text{n-gram}} \cap s_{\text{n-gram}} \right|} \tag{5}$$

where $|x_{\text{n-gram}} \cap s_{\text{n-gram}}|$ represents the number of overlapping n-grams between the source and the candidate sequence x for different values of $n \in \{1, \ldots, N\}$(we use $N = 3$). Since longer n-gram overlaps are more valuable, we set $\beta > 1$. This function inherently increases the BLEU score between the source and the generated sentences.

We address the semantic aspect of fidelity by devising a function based on the word embeddings of source and generated sentences. We define embedding based similarity between two sentences as,

$$\mathcal{S}(x, s) = \frac{1}{|x|} \sum_{w_i \in x} \underset{w_j \in s}{\text{argmax}}\ \psi(\mathbf{v}_{w_i}, \mathbf{v}_{w_j}) \tag{6}$$

where $\mathbf{v}_{w_i}$ is the word embedding for token w_i and $\psi(\mathbf{v}_{w_i}, \mathbf{v}_{w_j})$ is the gaussian radial basis function (rbf)[1]. For each word in the candidate sequence x, we find the best matching word in the source sentence using word level similarity. Using the above mentioned measure for embedding similarity we use the following submodular function:

$$\mathcal{L}_2(X, s) = \mu_2 \sqrt{\sum_{x \in X} \mathcal{S}(x, s)} \tag{7}$$

[1]We find gaussian rbf to work better than other similarity metrics such as cosine similarity

This function helps increase the semantic homogeneity between the source and generated sequences. The above defined functions (Equation 5,7) are compositions of non-decreasing concave functions and modular functions. Thus, staying in the realm of the class of monotone submodular functions mentioned in Equation 4, we define fidelity function $\mathcal{L}(X, s) = \mathcal{L}_1(X, s) + \mathcal{L}_2(X, s)$

Diversity

Ensuring high fidelity often comes at the cost of producing sequences that only slightly differ from each other. To encourage diversity in the generation process it is desirable to reward sequences with higher number of distinct n-grams as compared to others in the ground set $V^{(t)}$. Accordingly, we propose to use the following function:

$$\mathcal{D}_1(X) = \mu_3 \sum_{n=1}^{N} \beta^n \left| \bigcup_{x \in X} x_{n-gram} \right| \tag{8}$$

For $\beta = 1$, $\mathcal{D}_1(X)$ denotes the number of distinct n-grams present in the set X. Since shorter n-grams contribute more towards diversity, we set $\beta < 1$, thereby giving more value to shorter n-grams. It is easy to see that this function is monotone non-decreasing as the number of distinct n-grams can only increase with the addition of more sequences. To see that $\mathcal{D}_1(X)$ is submodular, consider adding a new sequence to two sets of sequences, one a subset of the other. Intuitively, the increment in the number of distinct n-grams when adding a new sequence to the smaller set should be larger than the increment when adding it to the larger set, as the distinct n-grams in the new sequence might have already been covered by the sequences in the larger set.

Apart from distinct n-gram overlaps, we also wish to obtain sequence candidates that are not only diverse, but also cover all major structural variations. It is reasonable to expect sentences that are structurally different to have lower degree of word/phrase alignment as compared to sentences with minor lexical variations. Edit distance (Levenshtein) is a widely accepted measure to determine such dissimilarities between two sentences. To incorporate this notion of diversity, a formulation in terms of edit distance seems like a natural fit for the problem. To do so, we use the coverage function which measures the similarity of the candidate sequences X with the ground set $V^{(t)}$. The

coverage function is naturally monotone submodular and is defined as:

$$\mathcal{D}_2(X) = \mu_4 \sum_{x_i \in V^{(t)}} \sum_{x_j \in X} \mathcal{R}(x_i, x_j) \quad (9)$$

where $\mathcal{R}(x_i, x_j)$ is an alignment based similarity measure between two sequences x_i and x_j given by:

$$\mathcal{R}(x_i, x_j) = 1 - \frac{\texttt{EditDistance}(x_i, x_j)}{|x_i| + |x_j|} \quad (10)$$

Note that $\mathcal{R}(x_i, x_j)$ will always lie in the range $[0, 1]$.

Evidently, this method allows flexibility in terms of controlling diversity and fidelity. Our goal is to strike a balance between these two to obtain high-quality generations.

5 Experiments

5.1 Datasets

In this section we outline the datasets used for evaluating our proposed method. We specify the actual splits in Table 2. Based on the task, we categorize them into the following:

1. **Intrinsic evaluation**: To demonstrate the efficacy of our method on fidelity and diversity, we use the standard *Quora question pair*[2] dataset and the *Twitter URL paraphrasing* (Lan et al., 2017) dataset.

   We train and evaluate the paraphrase generation model on a subset of *Quora question pair* dataset which we refer to as *Quora-Div*. This subset comprises only positive examples (pairs which have been annotated as paraphrases).

   We additionally perform in-domain data augmentation for the task of paraphrase recognition. For that, we augment sentences generated through different paraphrasing model as positive samples to the *Quora-PR* dataset. *Quora-PR* is a subset of *Quora question pair* dataset which contains positive as well as negative examples.

2. **Data-augmentation**: We exhibit the importance of samples generated through our method on the task of Data-augmentation using three primary datasets. *SNIPS (Coucke et al., 2018)*, *Yahoo-L31*[3] is used for intent-classification and *TREC* (Li and Roth, 2002) is used for question classification. Each dataset is balanced in terms of the number of samples per classes.

Dataset	Task	Train	Val.	Test	Classes
Quora-Div	Intrinsic	120K	20K	5K	N/A
Twitter	Intrinsic	100K	15K	3K	N/A
Quora-PR	Intrinsic	40K	10K	40K	2
DATA-AUGMENTATION					
SNIPS	Intent	10k	1k	700	7
Yahoo-L31	Intent	4K	1K	1K	2
TREC	Question	1K	200	500	6

Table 2: Dataset Statistics

5.2 Baseline

Several models have sought to increase diversity, albeit with different goals and techniques. However, majority of the prior works in this area have focused on the task of producing diverse responses in dialog systems (Li et al., 2016; Ritter et al., 2011) and not paraphrasing. Given the lack of relevant baselines, we compare our model against the following methods:

1. **SBS**: Decoder which performs standard beam search during generation.
2. **DBS**: An alternative of beam search to incorporate diversity. (Vijayakumar et al., 2018)
3. **DPP**: Decoder based on Determinantal Point Processes (Kulesza et al., 2012)
4. **SSR**[4]: Decoder based on Subset Selection using Simultaneous Sparse Recovery (Elhamifar et al., 2016)

   We additionally, evaluate against the following paraphrase generation models:

5. **VAE-SVG**: VAE based generative framework for paraphrase generation. (Gupta et al., 2018)
6. **RbM**: Deep Reinforcement learning based paraphrase generation model. (Li et al., 2018)

Note that the first four baselines are trained using the same SEQ2SEQ network and differ only in the decoding phase.

[2] https://www.kaggle.com/c/quora-question-pairs

[3] https://webscope.sandbox.yahoo.com/

[4] Exact formulation of the SSR and DPP can be found in the supplementary section.

Model	Quora-Div			Twitter		
	BLEU↑	METEOR↑	TERp↓	BLEU↑	METEOR↑	TERp↓
SBS	33.1	28.2	55.6	51.1	23.5	67.9
DBS (Vijayakumar et al., 2018)	30.9	28.3	57.5	47.1	22.1	69.0
VAE-SVG (Gupta et al., 2018)	33.4	25.6	63.2	46.7	25.2	67.1
RbM (Li et al., 2018)	29.4	29.5	62.5	47.7	29.3	68.7
DPP	30.5	27.9	57.3	44.8	21.4	71.4
SSR	28.7	26.8	58.7	41.3	20.0	74.4
DiPS (Ours)	**35.1**	**29.7**	**53.2**	**55.3**	**30.1**	**63.5**

Table 3: Results on **Quora-Div** and **Twitter** dataset. Higher↑ BLEU and METEOR score is better whereas lower↓ TERp score is better. Please see Section 6 for details.

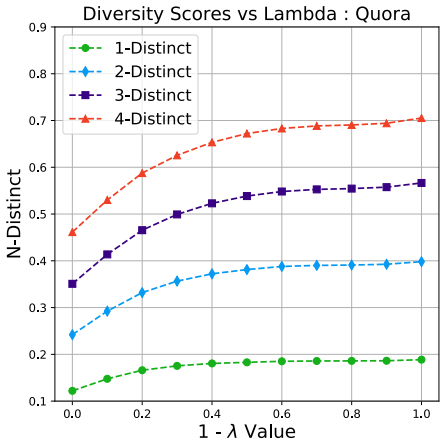

Figure 2: Effect of varying the trade-off coefficient λ in DiPS on various diversity metrics on the Quora dataset.

5.3 Intrinsic Evaluation

1. **Fidelity**: To evaluate our method for fidelity of generated paraphrases, we use three machine translation metrics which have been shown to be suitable for paraphrase evaluation task (Wubben et al., 2010): BLEU (Papineni et al., 2002)(upto bigrams), METEOR (Banerjee and Lavie, 2005) and TERPlus (Snover et al., 2009).
2. **Diversity**: We report degree of diversity by calculating the number of distinct n-grams (n $\in \{1, 2, 3, 4\}$). The value is scaled by the number of generated tokens to avoid favoring long sequences.

In addition to fidelity and diversity, we evaluate the efficacy of our method by using the generated paraphrases as augmented samples in the task of paraphrase recognition on the *Quora-PR* dataset. We perform experiments with multiple augmentation settings for the following classifiers:

1. **LogReg**: Simple Logistic Regression model. We use a set of hand-crafted features, the details of which can be found in the supplementary.
2. **SiameseLSTM**: Siamese adaptation of LSTM to measure quality between two sentences (Mueller and Thyagarajan, 2016)

We also perform ablation testing to highlight the importance of each submodular component. Details can be found in the supplementary section.

5.4 Data-Augmentation

We evaluate the importance of using high quality paraphrases in two downstream classification tasks namely intent-classification and question-classification. Our original generation model is trained on *Quora-Div* question pairs. Since intent-classification and question-classification contain questions, this setting seems like a good fit to perform transfer learning. We perform experiments on the following standard classifier models:

1. **LogRegDA**: Simple logistic regression model trained using hand-crafted features. For details, please refer to the supplementary section.
2. **LSTM**: Single layered LSTM classification model.

In addition to SBS and DBS, we use the following data-augmentation baselines for comparison:

1. **SynRep** : Simple synonym replacement
2. **ContAug**: Data-augmentation scheme based on replacement of words with similar paradigmatic relations. (Kobayashi, 2018)

5.5 Setup

We train our SEQ2SEQ model with attention (Bahdanau et al., 2014) for up to 50 epochs using the adam optimizer (Kingma and Ba, 2014) with initial learning rate set to 2e-4. During the generation phase, we follow standard beam search till

Model	Quora-Div				Twitter			
	1-distinct	2-distinct	3-distinct	4-distinct	1-distinct	2-distinct	3-distinct	4-distinct
SBS	12.8	24.8	35.3	46.6	20.0	30.9	38.1	44.6
VAE-SVG (Gupta et al., 2018)	15.8	22.5	27.6	31.8	19.3	28.2	33.3	37.2
DBS (Vijayakumar et al., 2018)	17.9	33.7	44.8	54.9	25.8	40.7	48.2	53.9
DPP	17.1	34.4	49.1	62.6	25.6	41.4	51.1	59.0
SSR	16.6	32.8	47.1	60.7	26.6	43.7	54.0	62.4
DiPS (Ours)	**18.1**	**37.2**	**52.3**	**65.3**	**28.3**	**46.6**	**56.7**	**64.5**

Table 4: Results on **Quora-Div** and **Twitter** dataset. Higher distinct scores imply better lexical diversity. Please see Section 6 for details.

Model	LogRegDA			LSTM		
	YahooL31	TREC	SNIPS	YahooL31	TREC	SNIPS
NoAug	62.7	82.2	93.4	64.8	94.2	94.7
SBS	63.6	84.6	93.8	65.4	94.4	94.7
DBS	63.3	84.2	94.1	65.6	95.2	96.1
SynRep	63.7	85.2	93.9	65.3	93.6	95.5
ContAug	63.8	86.0	95.3	66.3	95.8	96.4
DiPS(Ours)	**64.9**	**86.6**	**96.0**	**66.7**	**96.4**	**97.1**

Table 5: Accuracy scores of two classification models on various data-augmentation schemes. Please see Section 6 for details

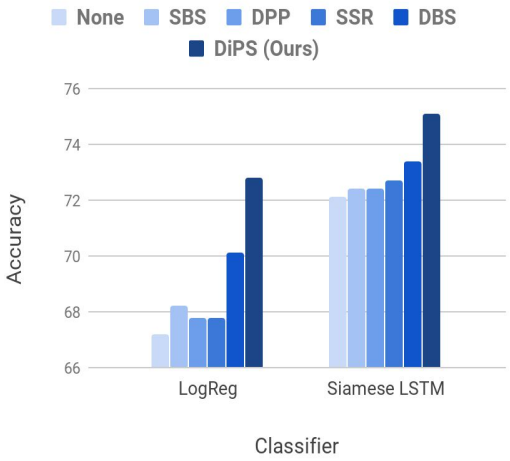

Figure 3: Comparison of accuracy scores of two paraphrase recognition models using different augmentation schemes (Quora-PR). Both LogReg and SiameseLSTM achieve the highest boost in performance when augmented with samples generated using DiPS

the number of generated tokens is nearly half the source sequence length (token level) to avoid possibly erroneous sentences. We then apply submodular maximization stochastically with probability p at each time step. Since each candidate subsequence is extended by a single token at every time-step, information added might not necessarily be useful as our submodular components work on sentence level. This approach is time efficient and avoids redundant computations.

For each augmentation setting, we randomly select sentences from the training data and generate its paraphrases. We then add them in the training data with the same label as that of the source sentence. We evaluate the performance on different classification models in terms of accuracy.

Based on the formulation of the objective function, it should be clear that diversity would attain maximum value at (or around) $\lambda = 0$ albeit at the cost of fidelity. This is certainly not a desirable property for paraphrasing systems. To address this, we perform hyperparameter tuning for λ value by analyzing the trade-off between diversity and fidelity based on varying λ values. In practice, diversity metric attains saturation at certain λ range (usually 0.2 - 0.5). This behaviour can be seen in Figure 2. Corresponding plot for Twitter, the effect of λ on fidelity and additional details about the hyperparameters can be found in the supplementary.

6 Results

Our experiments were geared towards answering the following primary questions:

Q1. Is DiPS able to generate diverse paraphrases without compromising on fidelity? (Section 6.1)

Q2. Are paraphrase generated by DiPS useful in data-augmentation? (Section 6.2)

6.1 Intrinsic Evaluation

We compare our method against recent paraphrasing models as well as multiple diversity inducing schemes. DiPS outperforms these baseline models in terms of fidelity metrics namely BLEU, METEOR and TERp. A high METEOR score and a low TERp score indicate the presence of not only exact words but also synonyms and semantically similar phrases. Notably, our model is not only able to achieve substantial gains over other diversity inducing schemes but is also able to do so

without compromising on fidelity. Diversity and fidelity scores are reported in Table 4 and Table 3, respectively.

As described in Section 5.3, we evaluate the accuracy of paraphrase recognition models when provided with training data augmented using different schemes. It is reasonable to expect that high quality paraphrases would tend to yield better results on in-domain paraphrase recognition task. We observe that using the paraphrases generated by DiPS helps in achieving substantial gains in accuracy over other baseline schemes. Figure 3 showcases the effect of using paraphrases generated by our method as compared to other competitive paraphrasing methods.

6.2 Data-augmentation

Data Augmentation results for intent and question classification are shown in Table 5. While, SBS does not offer much lexical variability, DBS offers high diversity at the cost of fidelity. SynRep and ContAug are augmentation schemes which are limited by the amount of structural variations they can offer. DiPS on the other hand provides generation having high structural variations without compromising on fidelity. The boost in accuracy scores on both the types of classification models is indicative of the importance of using high quality paraphrases for data-augmentation.

7 Conclusion

In this paper, we have proposed DiPS, a model which generates high quality paraphrases by maximizing a novel submodular objective function designed specifically for paraphrasing. In contrast to prior works which focus exclusively either on fidelity or diversity, a submodular function based approach offers a large degree of freedom to control fidelity and diversity. Through extensive experiments on multiple standard datasets, we have demonstrated the effectiveness of our approach over numerous baselines. We observe that the diverse paraphrases generated are not only interesting and meaning preserving, but are also helpful in data augmentation. We showcase that using multiple settings on the task of intent and question classification. We hope that our approach will be useful not only for paraphrase generation and data augmentation, but also for other NLG problems in conversational agents and text summarization.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. This work is supported in part by the Ministry of Human Research Development (Government of India), Amazon, and a travel gift from Microsoft Research (MSR) India.

References

Peter G Anick and Suresh Tipirneni. 1999. The paraphrase search assistant: terminological feedback for iterative information seeking. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 153–159. ACM.

Francis Bach et al. 2013. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends® in Machine Learning*, 6(2-3):145–373.

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Delphine Bernhard and Iryna Gurevych. 2008. Answering learners' questions by retrieving question paraphrases from social q&a sites. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 44–52. Association for Computational Linguistics.

Ziqiang Cao, Chuwei Luo, Wenjie Li, and Sujian Li. 2017. Joint copying and restricted generation for paraphrase.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190.

Mladen Dimovski, Claudiu Musat, Vladimir Ilievski, Andreea Hossman, and Michael Baeriswyl. 2018. Submodularity-inspired data selection for goal-oriented chatbot training based on sentence embeddings. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4019–4025. International Joint Conferences on Artificial Intelligence Organization.

Ehsan Elhamifar, Guillermo Sapiro, and S Shankar Sastry. 2016. Dissimilarity-based sparse subset selection. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2182–2197.

Ehsan Elhamifar, Guillermo Sapiro, and Rene Vidal. 2012. Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. In *Advances in Neural Information Processing Systems*, pages 19–27.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1608–1618.

Jenny Rose Finkel, Christopher D Manning, and Andrew Y Ng. 2006. Solving the problem of cascading errors: Approximate bayesian inference for linguistic annotation pipelines. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 618–626. Association for Computational Linguistics.

S Fujishige. 2005. Submodular functions and optimization. *Annals of Discrete Mathematics*, 58.

Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. 2013. A systematic exploration of diversity in machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *AAAI Conference on Artificial Intelligence*.

Rishabh K Iyer and Jeff A Bilmes. 2013. Submodular optimization with submodular cover and submodular knapsack constraints. In *Advances in Neural Information Processing Systems*, pages 2436–2444.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1875–1885.

Stefanie Jegelka and Jeff Bilmes. 2011. Submodularity beyond submodular energies: coupling edges in graph cuts.

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

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Katrin Kirchhoff and Jeff Bilmes. 2014. Submodularity for data selection in machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 131–141.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 452–457.

Vladimir Kolmogorov and Ramin Zabih. 2002. What energy functions can be minimized via graph cuts? In *European conference on computer vision*, pages 65–81. Springer.

Andreas Krause and Daniel Golovin. Submodular function maximization.

Andreas Krause and Carlos Guestrin. 2011. Submodularity and its applications in optimized information gathering. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(4):32.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Alex Kulesza, Ben Taskar, et al. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*.

Liangda Li, Ke Zhou, Gui-Rong Xue, Hongyuan Zha, and Yong Yu. 2009. Enhancing diversity, coverage and balance for summarization through structure learning. In *Proceedings of the 18th international conference on World wide web*, pages 71–80. ACM.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018. Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878. Association for Computational Linguistics.

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-Volume 1*, pages 510–520. Association for Computational Linguistics.

Kathleen R McKeown. 1983. Paraphrasing questions using given and new information. *Computational Linguistics*, 9(1):1–10.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity.

Preksha Nema, Mitesh M Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1063–1072.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functionsi. *Mathematical programming*, 14(1):265–294.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 85–91.

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.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255.

Aaditya Prakash, Sadid A Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual lstm networks. *arXiv preprint arXiv:1610.03098*.

Alexander J Ratner, Henry Ehrenberg, Zeshan Hussain, Jared Dunnmon, and Christopher Ré. 2017. Learning to compose domain-specific transformations for data augmentation. In *Advances in Neural Information Processing Systems*, pages 3239–3249.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593.

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)*, volume 1, pages 1073–1083.

Matthew Snover, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or hter?: exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268. Association for Computational Linguistics.

Yiping Song, Rui Yan, Yansong Feng, Yaoyuan Zhang, Dongyan Zhao, and Ming Zhang. 2018. Towards a neural conversation model with diversity net using determinantal point processes. In *AAAI*.

Peter Stobbe and Andreas Krause. 2010. Efficient minimization of decomposable submodular functions. In *Advances in Neural Information Processing Systems*, pages 2208–2216.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. AAAI Conference on Artificial Intelligence.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563.

Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer. 2010. Paraphrase generation as monolingual translation: Data and evaluation. In *INLGC*, pages 203–207. ACL.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association of Computational Linguistics*, 3(1):283–297.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.