Supplementary Material: 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

1 Introduction

In this supplementary paper we provide details about the model, additional experiments and formulation of the proposed baselines.

2 Model Details

2.1 Seq2Seq Models

Parameter	Value
Max grad norm	1.0
Batch size	16
Cell type	LSTM
LSTM Layers (Depth)	2
Hidden size	256
Embedding size	300
Vocabulary size	20,000
Dropout	None
Attention Model	Luong-general
Bidirectional Encoder	True
Max length	20
Learning Rate (Optimizer)	0.0002
Desired Paraphrases (k)	20

Table 1: SEQ2SEQ

Given a sequence of inputs $X = (x_1, \ldots, x_T)$, where T is the input sequence length, the goal of the sequence-to-sequence model is to estimate the conditional probability $\mathbb{P}(Y|X)$, where Y is the corresponding output sequence $Y = (y_1, \ldots, y_{T'})$. The input sequence length T may differ from the output sequence length T'. We choose the attention model (Luong et al., 2015; Bahdanau et al., 2014), which is based on the encoder-decoder framework proposed by (Cho et al., 2014; Sutskever et al., 2014). The encoder as well as the decoder is modeled using a recurrent neural network (RNN). We use a Long-short term memory unit (LSTM) (Hochreiter and Schmidhuber, 1997) as it helps in learning problems with long range temporal dependencies. The encoder LSTM takes as input the tokens of the sentence whose paraphrase needs to be generated and produces a sequence of encoder hidden states $h_i : i \in \{1 \ldots T\}$. At each time step, the decoder receives the word embedding of the previous word, a decoder state s_t and the attention distribution calculated using the weighted sum of encoder states:

$$c_t = \sum_{i=1}^{T} \alpha_{t_i} h_i, \;\; \alpha_{t_i} = \frac{\exp \eta(s_{t-1}, h_i)}{\sum_{j=1}^{T} \exp \eta(s_{t-1}, h_j)}$$

to produce the corresponding paraphrase token y'_t

2.2 Determinantal Point Processes (DPP)

Consider the problem of sampling S points from Y associated with a similarity matrix $K \in \mathbb{R}^{n \times n}$, that is symmetric, real and positive semi-definite (PSD). Determinantal point processes (DPP) (Kulesza et al., 2012) are elegant probabilistic models that capture negative correlation and help in efficient sampling which follow the distribution given by:

$$P(S \subseteq Y) = \det(K_S)$$

Assume the following q and ϕ functions:

$$q(x, s) = \frac{1}{|x|} \sum_{w_i \in x} \underset{w_j \in s}{\operatorname{argmax}} \; \psi(\mathbf{v}_{w_i}, \mathbf{v}_{w_j}) \qquad (1)$$

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

$$\phi(x_i, x_j) = \frac{1}{|x_i|} \sum_{w_k \in x_i} \underset{w_m \in x_j}{\text{argmax}}\ \psi(\mathbf{v}_{w_k}, \mathbf{v}_{w_m}) \tag{2}$$

Note that s is the source sentence, x_i, x_j are generated candidates. We calculate the kernel matrix, $L(x_i, x_j, s) = q(x_i, s)\phi(x_i, x_j)q(x_j)$ Note that this function is not symmetric. In order to make it symmetric we operate on the final kernel $K = \frac{1}{2}(L + L^\top)$

2.3 Subset selection via Simultaneous Sparse Recovery

Consider the problem of finding k points from a collection of $|V| = N$ data points which preserve the essential characteristics of the set $V = \{v_1, \ldots, v_N\}$. Assume that we can form a non-negative dissimilarity matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$ such that each element d_{ij} is indicative of how well a data point i is suited to be a representative of data point j. Elhamifar et al. (2012) propose a method to select a subset of points from V that can well encode all the data points based on the dissimilarity matrix $\mathbf{D}$.

To do so, consider variables $z_{ij} \in \mathbf{Z}$ associated with dissimilarities d_{ij}. Each element z_{ij} can be interpreted as the *probability* that data point i is a representative of j.They formulate the problem as the following row-sparsity regularized trace minimization program on $\mathbf{Z} \in \mathbb{R}^{N \times N}$:

$$\begin{aligned} &\min \text{tr}(\mathbf{D}^\top \mathbf{Z}) + \lambda \|\mathbf{Z}\|_{1,q} \\ &\text{s.t } \mathbf{Z} \geq \mathbf{0}, \mathbf{1}^\top \mathbf{Z} = \mathbf{1}^\top, \|\mathbf{Z}\|_{1,\infty} \leq k \end{aligned} \tag{3}$$

where k denotes the cardinality constraint, tr$(\cdot)$ denotes the trace operator, $\|Z\|_{1,q} \triangleq \sum_{i=1}^{N} \|z_i\|_q$ and $\mathbf{1}$ denotes an all-one N-dimensional vector. A set of representative points can be obtained by optimizing the above function and selecting indices corresponding to the non-zero rows of the sparse matrix $\mathbf{Z}^*$.

We start with selecting the top $3k$ most probable subsequences in each time step and then we use sparse subset selection to select k diverse subsequences which are fed into the decoder for the next time step. To use sparse subset selection we need to form a dissimilarity matrix $\mathbf{D}$. In contrast to DPP the matrix need not be positive semi-definite. In addition, elements d_{ij}, need not necessarily satisfy triangle inequality and the matrix $\mathbf{D}$ can be asymmetric as well. We use an alternate formulation of Sparse subset selection (Elhamifar et al., 2016) to select k-samples from a given ground set:

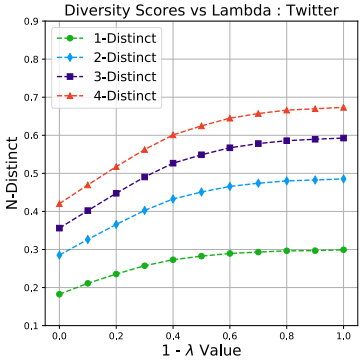

Figure 1: Effect of varying the trade-off coefficient λ in DiPS on various diversity metrics.

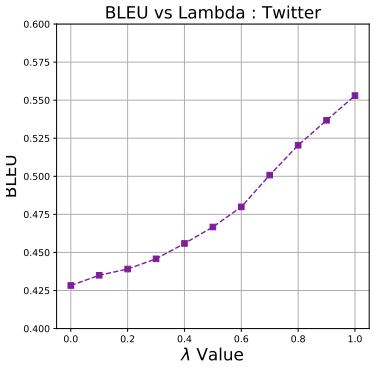

Figure 2: Effect of varying the trade-off coefficient λ in DiPS on BLEU score for twitter dataset.

$$\begin{aligned} &\min \text{tr}(\mathbf{D}^\top \mathbf{Z}) \\ &\text{s.t } \|\mathbf{Z}\|_{1,\infty} \leq k,\ \mathbf{Z} \geq \mathbf{0}, \mathbf{1}^\top \mathbf{Z} = \mathbf{1}^\top, \end{aligned} \tag{4}$$

We use the following equation to compute dissimilarity between two sequences:

$$\mathbf{D}_{ij} = 1 - \phi(x_i, x_j)$$

3 Experiments: Ablation

In this section, we highlight the importance of using each submodular component towards generation of high quality paraphrases.

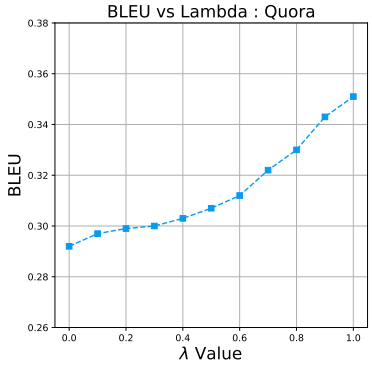

Figure 3: Effect of varying the trade-off coefficient λ in DiPS on BLEU score for quora dataset.

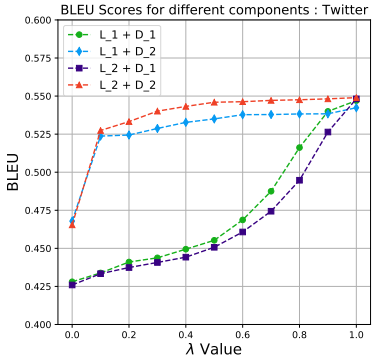

Figure 4: Effect of varying the trade-off coefficient λ in DiPS for individual combinations of submodular components.

Submodular Components	**BLEU**	**2-distinct**
$\mathcal{L}_1 + \mathcal{D}_1$	48.7	48.0
$\mathcal{L}_1 + \mathcal{D}_2$	52.3	35.4
$\mathcal{L}_2 + \mathcal{D}_1$	46.0	46.5
$\mathcal{L}_2 + \mathcal{D}_2$	51.6	35.5

Table 2: Results of ablataion testing

References

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

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

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.