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

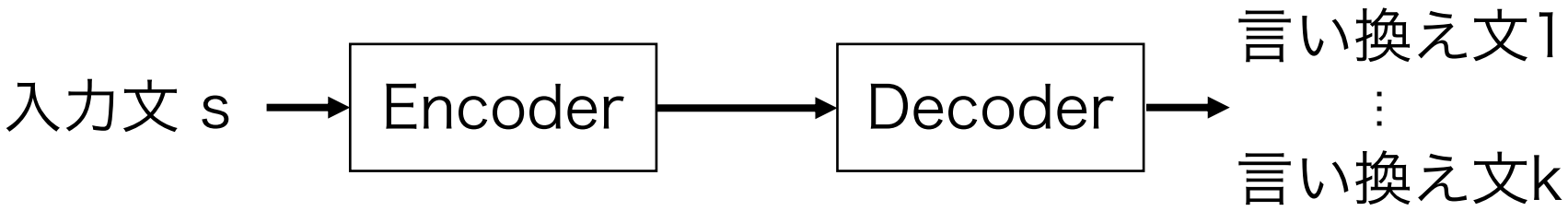
**Ashutosh Kumar**[*1] **Satwik Bhattamishra**[*2 †] **Manik Bhandari**[1] **Partha Talukdar**[1]
[1] Indian Institute of Science, Bangalore, India
[2] Birla Institute of Technology and Science, Pilani, India
{ashutosh, ppt}@iisc.ac.in, {satwik55, mbbhandarimanik}@gmail.com

読む人：今野 颯人

乾・鈴木研究室 M1

# 概要

- 入力文sからEnc-Decによりk個の言い換え文を得る




- 応用先：Data augmentation, 対話エージェント
- 目的：従来より多様性のある言い換え文を得たい
- 提案手法：劣モジュラ関数最大化として定式化
- 評価：従来よりも元の文に忠実かつ多様な文を
  生成することに成功
  Data Augmentationでの効果も検証

# 実際の入出力

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

従来よりも多様

# モチベーション

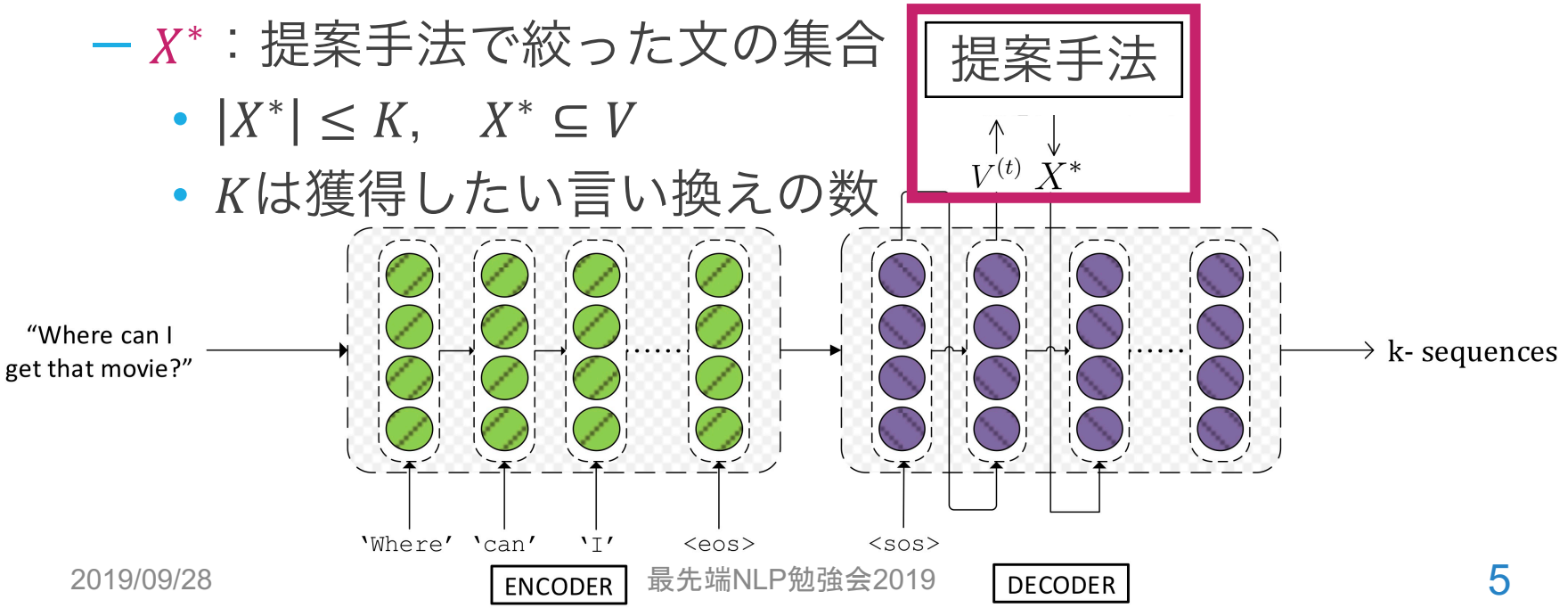
- 言い換えの質を決める2つの要素：Fidelity，Diversity
  - Fidelity：元の文に忠実か（意味的類似性）
  - Diversity：語彙的にどのくらい異なるか（多様性）

- 従来の手法：top-k beam search
  - 構文が類似，単語レベルの変化のみ，
  - Diversityが低い

- 目的：Fidelityを落とさずにDiversityを向上させたい
- 提案手法：Fidelity，Diversityを最大化するために
  劣モジュラ関数最大化として定式化

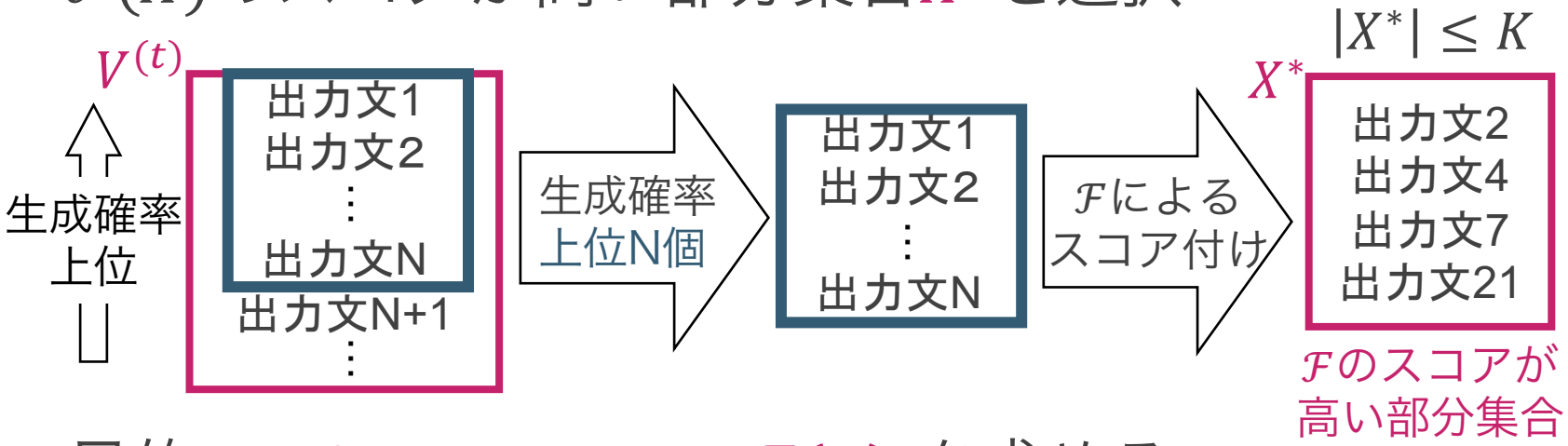
# 提案手法

- まずはSEQ2SEQをcross entropy lossで学習
- デコード時のtime-step t ごとに
  出力文の集合$V^{(t)}$ からFidelityとDiversityが高い
  文の部分集合$X^*$を選ぶ
  - $V^{(t)}$：t時点での出力文(確率付き)の集合
  - $X^*$：提案手法で絞った文の集合
    - $|X^*| \leq K, \quad X^* \subseteq V$
    - $K$は獲得したい言い換えの数



# 提案手法

- $V^{(t)}$から$X^*$へ絞りたい
- 集合関数 $\mathcal{F}：2^V \rightarrow \mathbb{R}$ を作成
  - $\mathcal{F}(X)$： 部分集合$X$に対してスコア付けする関数
  - $\mathcal{F}(X)$のスコアが高い=文の部分集合$X$の FidelityとDiversityが高い(ように集合関数$\mathcal{F}$を作成)
- $\mathcal{F}(X)$のスコアが高い部分集合$X^*$を選択




- 目的： $X^* = \mathrm{argmax}_{X \subseteq V} \mathcal{F}(X)$ を求める

# 組み合わせ爆発の問題

- $X^* = \mathrm{argmax}_{X \subseteq V} \mathcal{F}(X)$ を求めたい
- しかし$\mathcal{F}(X)$を最大化する部分集合$X^*$を選ぶのは困難
  - 冪集合$2^V$から最適解となる部分集合を選ぶ
  - NP困難な最適化問題
- 台集合の要素数や$|X|$の上限値$K$が大きくなると
  解の候補は爆発的に大きくなる
- $\mathcal{F}$が非負で非単調減少の劣モジュラ関数ならば
  貪欲法で最適解の63.2%$(1 - 1/e)$に近似できる

# 貪欲法による近似解

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

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

$\mathcal{F}$は非負で非単調減少の劣モジュラ関数

```
1 S ← ∅
2 N ← V
3 while |S| < k do
4 |   x* ← argmax_{x∈N} F(S ∪ {x})
5 |   S ← S ∪ {x*}
6 |   N ← N \ {x*}
7 end
8 return S
```

$S = \{\}$からスタートし
$|S| = k$となるまで
「貪欲」に要素を
1つずつ増やしていく

$X^* = \mathrm{argmax}_{X \subseteq V} \mathcal{F}(X)$の近似解を得る

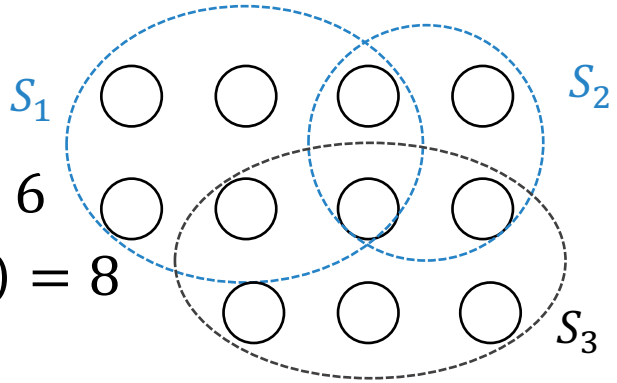
# 劣モジュラ関数

- 劣モジュラ関数：劣モジュラ性を持った集合関数
- 劣モジュラ性

$$\forall A \subseteq B \subseteq V,\ \ \forall j \in V \setminus B:$$
$$\mathcal{F}(A \cup \{j\}) - \mathcal{F}(A) \geq \mathcal{F}(B \cup \{j\}) - \mathcal{F}(B)$$

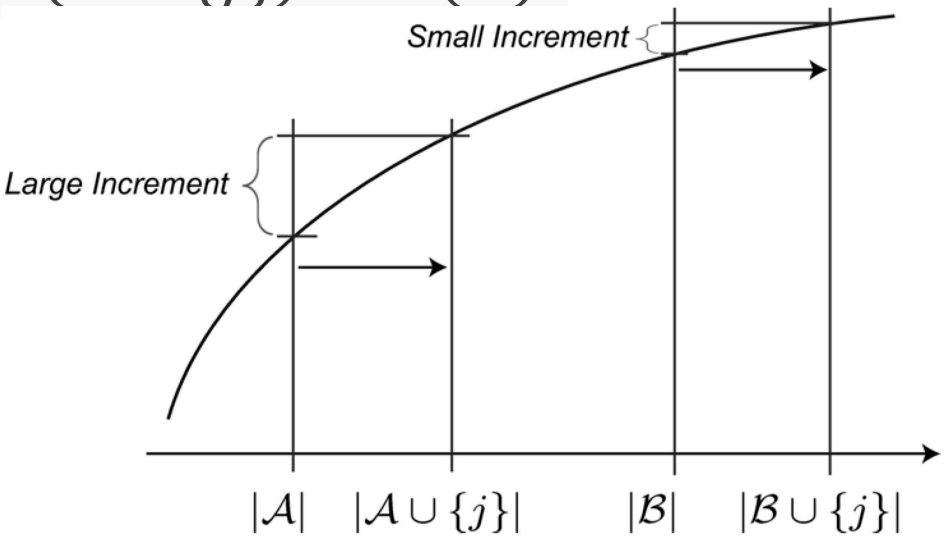
サイズとともに
増加が緩やかとなる

例）カバー関数




$f(\{S_1\}) = 6$

$f(\{S_1, S_2\}) = 8$




$$f(\{S_1, S_3\}) - f(\{S_1\}) > f(\{S_1, S_2, S_3\}) - f(\{S_1, S_2\})$$
$$10 - 6 = 4 \qquad\qquad 11 - 8 = 3$$

参考文献：河原吉伸，機械学習における劣モジュラ性の利用と組合せ論的アルゴリズム，オペレーションズ・リサーチ2013年5号

# 劣モジュラ関数の作成

- 文の部分集合$X$のFidelityとDiversityが高いと$\mathcal{F}(X)$のスコアが高いように集合関数$\mathcal{F}$を作成
- $\mathcal{F}(X) = \lambda\mathcal{L}(X,s) + (1-\lambda)\mathcal{D}(X)$
  - $\mathcal{L}(X,s)$： Fidelityの良さを表す集合関数
    $$\mathcal{L}(X,s) = \mathcal{L}_1(X,s) + \mathcal{L}_2(X,s)$$
  - $\mathcal{D}(X)$： Diversityの良さを表す集合関数
    $$\mathcal{D}(X) = \mathcal{D}_1(X) + \mathcal{D}_2(X)$$
    - $s$：入力文，$\lambda \in [0,1]$
- $\mathcal{L}_1(X,s), \mathcal{L}_2(X,s), \mathcal{D}_1(X), \mathcal{D}_2(X)$は全て非負で非単調減少の劣モジュラ関数

# 劣モジュラ関数の作成①：Fidelity

- $\mathcal{F}(X) = \lambda \mathcal{L}(X, s) + (1 - \lambda)\mathcal{D}(X)$
- Fidelity：$\mathcal{L}(X, s) = \mathcal{L}_1(X, s) + \mathcal{L}_2(X, s)$

- $\mathcal{L}_1(X, s) = \mu_1 \sqrt{\sum_{x \in X} \sum_{n=1}^{N} \beta^n \left| x_{n-gram} \cap s_{n-gram} \right|}$
  - 元の文と言い換えのN-grams重複度

- $\mathcal{L}_2(X, s) = \mu_2 \sqrt{\sum_{x \in X} \frac{1}{|x|} \sum_{w_i \in x} \operatorname{argmax} \psi \left( v_{w_i}, v_{w_j} \right)}$
  - 元の文と言い換えの単語レベル類似度

# 劣モジュラ関数の作成②：Diversity

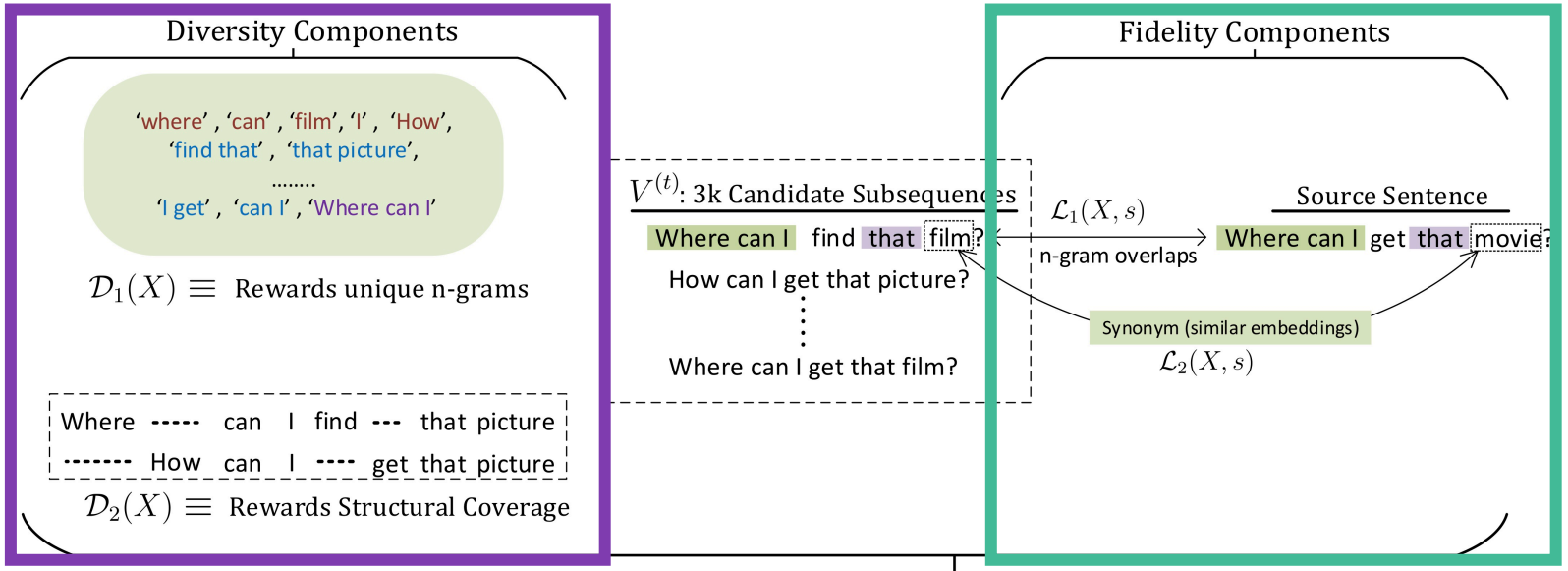
- $\mathcal{F}(X) = \lambda \mathcal{L}(X, s) + (1 - \lambda)\mathcal{D}(X)$
- Diversity：$\mathcal{D}(X) = \mathcal{D}_1(X) + \mathcal{D}_2(X)$

- $\mathcal{D}_1(X) = \mu_3 \sum_{n=1}^{N} \beta^n \left| \cup_{x \in X} x_{n-gram} \right|$
  - 部分集合$X$でのN-gramsの異なり度合い

- $\mathcal{D}_2(X) = \mu_4 \sum_{x_i \in V^{(t)}} \sum_{x_j \in X} \left( 1 - \frac{EditDistance(x_i, x_j)}{|x_i| + |x_j|} \right)$
  - 部分集合$X$での編集距離

# 提案手法

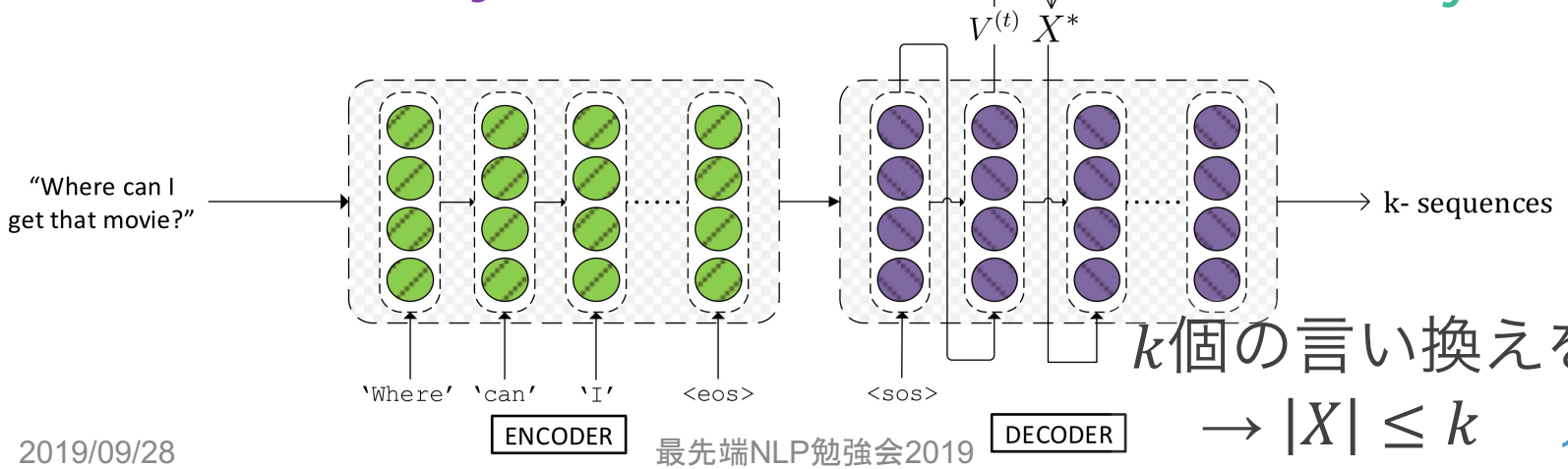
$$\mathcal{F}(X) = (1-\lambda)\mathcal{D}(X) + \lambda\mathcal{L}(X,s)$$




Diversity

Fidelity

$$argmax_{X \subset V^{(t)}} \mathcal{F}(X)$$




$k$個の言い換えを得る

$\rightarrow |X| \leq k$

# 検証

- 1. 言い換え文のQualityは良いか？
  - Fidelityを損なうことなくDiversityを向上させられるか
- 2. Data Augmentationとして有効か？
  - ２つのタスクで検証
    - Intent-Classification
    - Question Classification

# 検証①言い換えのQuality：Diversity

- 評価指標：N-distinct
- Dataset：
  - Quora-question pair dataset
  - Twitter URL paraphrasing dataset

| Model | Quora-Div | | | |
|---|---|---|---|---|
| | 1-distinct | 2-distinct | 3-distinct | 4-distinct |
| SBS | 12.8 | 24.8 | 35.3 | 46.6 |
| VAE-SVG (Gupta et al., 2018) | 15.8 | 22.5 | 27.6 | 31.8 |
| DBS (Vijayakumar et al., 2018) | 17.9 | 33.7 | 44.8 | 54.9 |
| DPP | 17.1 | 34.4 | 49.1 | 62.6 |
| SSR | 16.6 | 32.8 | 47.1 | 60.7 |
| DiPS (Ours) | **18.1** | **37.2** | **52.3** | **65.3** |

従来よりも多様性のある文を生成

# 検証①言い換えのQuality：Fidelity

- 評価指標：BLUE, METEOR, TER-plus
- Dataset：
  - Quora-question pair dataset
  - Twitter URL paraphrasing dataset

| Model | Quora-Div | | |
|---|---|---|---|
| | BLEU↑ | METEOR↑ | TERp↓ |
| SBS | 33.1 | 28.2 | 55.6 |
| DBS (Vijayakumar et al., 2018) | 30.9 | 28.3 | 57.5 |
| VAE-SVG (Gupta et al., 2018) | 33.4 | 25.6 | 63.2 |
| RbM (Li et al., 2018) | 29.4 | 29.5 | 62.5 |
| DPP | 30.5 | 27.9 | 57.3 |
| SSR | 28.7 | 26.8 | 58.7 |
| DiPS (Ours) | **35.1** | **29.7** | **53.2** |

Fidelityを損なわずに文を生成

# 検証②Data Augmentation:

- Transfer Learningで分類問題のData Augmentationとして使用
  - 言い換え文生成モデルはQuora-Div question pairsで学習
- Intent-classification：YahooL31，SNIPS
- Question Classification：TREC

| 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** |

Data Augmentationにより精度向上

# まとめ

- 言い換え文のクオリティに重要なFidelityとDiversityについて劣モジュラ関数最大化として定式化
- 従来の手法よりも高品質な言い換え文を生成
- Data-AugmentationによりIntent-ClassificationとQuestion Classificationで有用性が示された
- Codeも公開済み：
  - https://github.com/malllabiisc/DiPS

# 感想

- 近似しているにも関わらずdecodeに時間がかかりそう
- Data Augmentationで有効であるかはTransfer Learning次第？
  - 今回の実験設定ではIntent-classificationとQuestion-classificationにquestionが含まれているため，question pairsで学習したのが上手くいった