Diverse Paraphrasing and its Effectiveness in Data Augmentation

Ashutosh Kumar*, Satwik Bhattamishra*, Manik Bhandari, Partha Talukdar
Machine and Language Learning Lab (MALL)
Indian Institute of Science, Bangalore

*Equal Contributions
Conversational Agents
Conversational Agents
Conversational Agents

User: I want to book a flight from Minneapolis to New York.
Conversational Agents

User: I want to book a flight from Minneapolis to New York

Bot: Sure. When are you planning to travel?
Conversational Agents

User: I want to book a flight from Minneapolis to New York

Bot: Sure. When are you planning to travel?

User: Can you book plane to New York from Minneapolis
Conversational Agents

Sorry, I don’t understand what you’re saying

User: I want to book a flight from Minneapolis to New York
Bot: Sure. When are you planning to travel?
User: Can you book plane to New York from Minneapolis
Bot: Sorry, I don’t understand what you’re saying
Conversational Agents

Data augmentation might help

User: I want to book a flight from Minneapolis to New York

Bot: Sure. When are you planning to travel?

User: Can you book plane to New York from Minneapolis

Bot: Sorry, I don’t understand what you’re saying

Bot: Sorry, I don’t understand what you’re saying
Paraphrase Generation

Rephrasing a given text in multiple ways
Paraphrase Generation

Rephrasing a given text in multiple ways

| Source          | how do i increase body height? |
# Paraphrase Generation

**Rephrasing a given text in multiple ways**

<table>
<thead>
<tr>
<th>Source</th>
<th>how do i increase body height ?</th>
</tr>
</thead>
</table>
| Paraphrases | • how could i increase my height ?  
• what should i do to increase body height ?  
• what are the ways to increase height ?  
• are there some ways to increase body height ? |
## Paraphrase Generation

Rephrasing a given text in multiple ways

<table>
<thead>
<tr>
<th>Source</th>
<th>how do i increase body height?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paraphrases</td>
<td>how could i increase my height?</td>
</tr>
<tr>
<td></td>
<td>what should i do to increase body height?</td>
</tr>
<tr>
<td></td>
<td>what are the ways to increase height?</td>
</tr>
<tr>
<td></td>
<td>are there some ways to increase body height?</td>
</tr>
</tbody>
</table>

**Fidelity**

(Meaning preserving)
### Paraphrase Generation

**Rephrasing a given text in multiple ways**

<table>
<thead>
<tr>
<th>Source</th>
<th>how do i increase body height?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Paraphrases</strong></td>
<td>how could i increase my height?</td>
</tr>
<tr>
<td></td>
<td>what should i do to increase body height?</td>
</tr>
<tr>
<td></td>
<td>what are the ways to increase height?</td>
</tr>
<tr>
<td></td>
<td>are there some ways to increase body height?</td>
</tr>
</tbody>
</table>

**Fidelity**

*Meaning preserving*

**Diversity**

*Lexical & syntactical variety*
Current State
Current State

Synonym or phrase replacement
<table>
<thead>
<tr>
<th>Synonym or phrase replacement</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>how do i increase body height ?</td>
<td></td>
</tr>
<tr>
<td>Synonym or phrase replacement</td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>how do i increase body height ?</td>
</tr>
<tr>
<td><strong>Synonym</strong></td>
<td>how do i <strong>grow</strong> body height ?</td>
</tr>
</tbody>
</table>
## Current State

<table>
<thead>
<tr>
<th>Synonym or phrase replacement</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>how do i increase body height ?</td>
</tr>
<tr>
<td>Synonym</td>
<td>how do i grow body height ?</td>
</tr>
<tr>
<td>Phrase</td>
<td>how do i increase the body measurement vertically?</td>
</tr>
</tbody>
</table>
Current State

<table>
<thead>
<tr>
<th>Synonym or phrase replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
</tr>
<tr>
<td>how do i increase body height ?</td>
</tr>
<tr>
<td>Synonym</td>
</tr>
<tr>
<td>how do i grow body height ?</td>
</tr>
<tr>
<td>Phrase</td>
</tr>
<tr>
<td>how do i increase the body measurement vertically?</td>
</tr>
</tbody>
</table>

Sentence → Encoder → Decoder → Paraphrase
Current State

<table>
<thead>
<tr>
<th>Source</th>
<th>how do i increase body height?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonym</td>
<td>how do i grow body height?</td>
</tr>
<tr>
<td>Phrase</td>
<td>how do i increase the body measurement vertically?</td>
</tr>
</tbody>
</table>

Synonym or phrase replacement

Sentence → Encoder → Decoder → Paraphrase

Subsequence Selection - Beam Search (Top-k)
Current State

### Synonym or phrase replacement

<table>
<thead>
<tr>
<th>Source</th>
<th>how do i increase body height?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonym</td>
<td>how do i grow body height?</td>
</tr>
<tr>
<td>Phrase</td>
<td>how do i increase the body measurement vertically?</td>
</tr>
</tbody>
</table>

### Subsequence Selection - Beam Search (Top-k)

<table>
<thead>
<tr>
<th>Source</th>
<th>how do i increase body height?</th>
</tr>
</thead>
</table>
# Current State

## Synonym or phrase replacement

<table>
<thead>
<tr>
<th>Source</th>
<th>how do i increase body height?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonym</td>
<td>how do i grow body height?</td>
</tr>
<tr>
<td>Phrase</td>
<td>how do i increase the body measurement vertically?</td>
</tr>
</tbody>
</table>

## Subsequence Selection - Beam Search (Top-k)

<table>
<thead>
<tr>
<th>Source</th>
<th>how do i increase body height?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beam</td>
<td>• how do i increase my height?</td>
</tr>
<tr>
<td></td>
<td>• how do i increase my body height?</td>
</tr>
<tr>
<td></td>
<td>• how do i increase the height?</td>
</tr>
<tr>
<td></td>
<td>• how would i increase my body height?</td>
</tr>
</tbody>
</table>
Current State

<table>
<thead>
<tr>
<th>Synonym or phrase replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source</strong></td>
</tr>
<tr>
<td>how do i increase body height ?</td>
</tr>
<tr>
<td><strong>Synonym</strong></td>
</tr>
<tr>
<td>how do i grow body height ?</td>
</tr>
<tr>
<td><strong>Phrase</strong></td>
</tr>
<tr>
<td>how do i increase the body measurement vertically ?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subsequence Selection - Beam Search (Top-k)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source</strong></td>
</tr>
<tr>
<td>how do i increase body height ?</td>
</tr>
<tr>
<td><strong>Beam</strong></td>
</tr>
<tr>
<td>• how do i increase my height ?</td>
</tr>
<tr>
<td>• how do i increase my body height ?</td>
</tr>
<tr>
<td>• how do i increase the height ?</td>
</tr>
<tr>
<td>• how would i increase my body height ?</td>
</tr>
</tbody>
</table>

Fidelity ✔
Current State

Synonym or phrase replacement

<table>
<thead>
<tr>
<th>Source</th>
<th>how do i increase body height?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonym</td>
<td>how do i grow body height?</td>
</tr>
<tr>
<td>Phrase</td>
<td>how do i increase the body measurement vertically?</td>
</tr>
</tbody>
</table>

Subsequence Selection - Beam Search (Top-k)

<table>
<thead>
<tr>
<th>Source</th>
<th>how do i increase body height?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beam</td>
<td>how do i increase my height?</td>
</tr>
<tr>
<td></td>
<td>how do i increase my body height?</td>
</tr>
<tr>
<td></td>
<td>how do i increase the height?</td>
</tr>
<tr>
<td></td>
<td>how would i increase my body height?</td>
</tr>
</tbody>
</table>

Fidelity ✅ Diversity ✗
What can we do?
What can we do?

Subsequence Selection - Beam Search (Diverse selection)
What can we do?

<table>
<thead>
<tr>
<th>Subsequence Selection - Beam Search (Diverse selection)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
</tr>
</tbody>
</table>
What can we do?

<table>
<thead>
<tr>
<th>Source</th>
<th>how do i increase body height?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beam</td>
<td>• how do i increase my height?</td>
</tr>
<tr>
<td></td>
<td>• how can i decrease my body weight?</td>
</tr>
<tr>
<td></td>
<td>• what do i do to increase the height?</td>
</tr>
<tr>
<td></td>
<td>• i am 17, what steps to take to decrease weight?</td>
</tr>
</tbody>
</table>
What can we do?

<table>
<thead>
<tr>
<th>Subsequence Selection - Beam Search (Diverse selection)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source</strong></td>
</tr>
<tr>
<td>Beam</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Diversity ✅
## What can we do?

<table>
<thead>
<tr>
<th>Subsequence Selection - Beam Search (Diverse selection)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source</strong></td>
</tr>
<tr>
<td>Beam</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**Fidelity ✗**  **Diversity ✓**
What we need
What we need

Fidelity ✓ Diversity ✓
What we need

Fidelity ✓ Diversity ✓

DiPS
Find $k$ diverse paraphrases with high fidelity
Method based on subset selection of candidate (sub)sequences
<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>how do i increase my ...</td>
<td></td>
</tr>
<tr>
<td>how can i decrease the ...</td>
<td></td>
</tr>
<tr>
<td>how can i grow the ...</td>
<td></td>
</tr>
<tr>
<td>what ways exist to increase ...</td>
<td></td>
</tr>
<tr>
<td>how would I increase the ...</td>
<td></td>
</tr>
<tr>
<td>how do I decrease the ...</td>
<td></td>
</tr>
<tr>
<td>i am 17 , what ...</td>
<td></td>
</tr>
<tr>
<td>are there ways to increase ...</td>
<td></td>
</tr>
</tbody>
</table>

$V^t$
Subset Selection

V^t \rightarrow X
Subset Selection

\[ \text{argmax}_{X \subseteq V_t, |X| = k} F(X) \]

- how do i increase my ...
- how can i decrease the ...
- how can i grow the ...
- what ways exist to increase ...
- how would I increase the ...
- how do I decrease the ...
- i am 17, what ...
- are there ways to increase ...

\[ V_t \rightarrow X \]

\[ k \]
Subset Selection

If $F$ is sub modular + monotone = Greedy algo. with good bounds exists

\[
\text{argmax}_{X \subseteq V, |X|=k} F(X)
\]
Sub-modularity

\[ F = \# \text{ Unique Coloured items} \]
Sub-modularity

F = # Unique Coloured items

# Items = 4
F = 2
Sub-modularity

\[ F = \# \text{Unique Coloured items} \]

# Items = 4
\[ F = 2 \]
Sub-modularity

\[ F = \# \text{ Unique Coloured items} \]

# Items = 4
\[ F = 2 \]

# Items = 4 + 1
\[ F = 2 + 1 \]
Sub-modularity

F = # Unique Coloured items

# Items = 4
F = 2

# Items = 4 + 1
F = 2 + 1
Sub-modularity

\[ F = \# \text{Unique Coloured items} \]

# Items = 4
\[ F = 2 \]

# Items = 4 + 1
\[ F = 2 + 1 \]

# Items = 5 + 1
\[ F = 3 + 0 \]
Sub-modularity

\[ F = \# \text{Unique Coloured items} \]

# Items = 4
\[ F = 2 \]

# Items = 4 + 1
\[ F = 2 + 1 \]

# Items = 5 + 1
\[ F = 3 + 0 \]
Monotonicity

\( F(A) \leq F(B) \)
DiPS

Induce Diversity while not compromising on Fidelity

\[ \mathcal{D}_1(X) \equiv \text{Rewards unique n-grams} \]

\[ \mathcal{D}_2(X) \equiv \text{Rewards Structural Coverage} \]

\[ \arg\max_{X \subset V(t)} F(X) \]

\[ V(t) \text{; 3k Candidate Subsequences} \]

\[ \mathcal{L}_1(X, s) \rightarrow \text{Source Sentence} \]

\[ \mathcal{L}_2(X, s) \]

\[ \text{Encoder} \]

\[ \text{Decoder} \rightarrow k\text{-sequences} \]
DiPS
Induce Diversity while not compromising on Fidelity

Diversity Components


‘I get’, ‘can I’, ‘Where can I’

\( D_1(X) \equiv \text{Rewards unique n-grams} \)

\( D_2(X) \equiv \text{Rewards Structural Coverage} \)

Fidelity Components

\( V(t): 3k \text{ Candidate Subsequences} \)

\( \mathcal{L}_1(X, s) \text{ Source Sentence} \)

\( \mathcal{L}_2(X, s) \text{ Synonym (similar embeddings)} \)

\( \arg\max_{X \subset V(t)} \mathcal{F}(X) \)

"Where can I get that movie?"

"Where can I find that picture?"

"How can I get that picture?"

Encoder

Decoder

\( \text{ENCODER} \)

\( \text{DECODER} \)

\( \text{k-sequences} \)
DiPS

Induce Diversity while not compromising on Fidelity

\[ \mathcal{D}_1(X) \equiv \text{Rewards unique n-grams} \]

\[ \mathcal{D}_2(X) \equiv \text{Rewards Structural Coverage} \]

Source Sentence

Where can I find that film?

Where can I get that picture?

Where can I get that movie?

**Encoder**

```
<where> 'can' 'I' <eos>
```

**Decoder**

```
<eos> 'I' 'can' 'where' 
```

\[ \text{V}^{(t)}; 3k \text{ Candidate Subsequences} \]

\[ \mathcal{L}_1(X, s) \]

\[ \mathcal{L}_2(X, s) \]

\[ \text{argmax}_{X \subseteq V^{(t)}} \mathcal{F}(X) \]
DiPS

Induce Diversity while not compromising on Fidelity

Diversity Components


\[ D_2(X) \equiv \text{Rewards Structural Coverage} \]

Fidelity Components

\[ V^{(t)}: 3k \text{ Candidate Subsequences} \]

\[ L_1(X, s) \rightarrow \text{Source Sentence} \rightarrow L_2(X, s) \]

\[ \text{Synonym (similar embeddings)} \]

Diversity

Encoder

Decoder

\[ \arg\max_{X \subseteq V^{(t)}} F(X) \]

argmax and V^{(t)}

Where can I find that film?

How can I get that picture?

Where can I get that film?

argmax and X^*

Where can I get that movie?
DiPS

Induce Diversity while not compromising on Fidelity

Diversity Components

- 'where', 'can', 'film', 'I', 'How', 'find that', 'that picture', ...
- 'I get', 'can I', 'Where can I'

Diversity

Fidelity Components

Source Sentence

- Where can I get that movie?
- How can I get that picture?
- Where can I get that film?

$V^{(t)}$: 3k Candidate Subsequences

$\mathcal{L}_1(X, s)$

Synonym (similar embeddings)

$\mathcal{L}_2(X, s)$

$D_2(X) \equiv$ Rewards Structural Coverage

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

"Where can I get that movie?"

$k$-sequences

Encoder

Decoder

ENCODER

DECODER
DiPS

Induce Diversity while not compromising on Fidelity

Diversity Components

\[ D_1(X) \equiv \text{Rewards unique n-grams} \]

\[ D_2(X) \equiv \text{Rewards Structural Coverage} \]

Fidelity Components

\[ V^{(t)}: 3k \text{ Candidate Subsequences} \]

\[ L_1(X, s) \rightarrow \text{Source Sentence} \]

\[ L_2(X, s) \rightarrow \text{Synonym (similar embeddings)} \]

\[ \arg \max_{X \subset V^{(t)}} \mathcal{F}(X) \]

\[ k \text{- sequences} \]
DiPS

Induce Diversity while not compromising on Fidelity

Diversity Components

\[ D_1(X) \equiv \text{Rewards unique n-grams} \]

\[ \text{\textit{where}}, \text{\textit{can}}, \text{\textit{film}}, \text{\textit{I}}, \text{\textit{How}}, \]

\[ \text{\textit{find that}}, \text{\textit{that picture}}, \]

\[ \text{\textit{i get}}, \text{\textit{can I}}, \text{\textit{Where can I}} \]

Fidelity Components

\[ V^{(t)}; \text{3k Candidate Subsequences} \]

\[ \mathcal{L}_1(X, s) \]

Source Sentence

\[ \text{Where can I get \textit{that} movie?} \]

\[ \text{Synonym (similar embeddings) } \]

\[ \mathcal{L}_2(X, s) \]

\[ \text{argmax}_{X \subseteq V^{(t)}} \mathcal{F}(X) \]

ENCODER

DECODER

“Where can I get that movie?”

\[ V^{(t)} \]

\[ X^* \]

\[ k\text{- sequences} \]
DiPS
Induce Diversity while not compromising on Fidelity

\[ \mathcal{D}_1(X) \equiv \text{Rewards unique n-grams} \]

\[ \mathcal{D}_2(X) \equiv \text{Rewards Structural Coverage} \]

\[ V^{(t)}: 3k \text{ Candidate Subsequences} \]

\[ L_1(X, s) \]

\[ \text{Source Sentence} \]

\[ \text{Synonym (similar embeddings)} \]

\[ L_2(X, s) \]

\[ \arg\max_{X \subset V^{(t)}} \mathcal{F}(X) \]

\[ \to k- \text{sequences} \]

“Where can I get that movie?”


- Rewards unique n-grams

- Rewards Structural Coverage

- Where can I find that picture

- How can I get that picture

- Rewards unique n-grams

- Rewards Structural Coverage
DiPS

Induce Diversity while not compromising on Fidelity

\[ \mathcal{D}_1(X) \equiv \text{Rewards unique n-grams} \]

\[ \mathcal{D}_2(X) \equiv \text{Rewards Structural Coverage} \]

**Diversity Components**


**Fidelity Components**

\[ V^{(t)}; 3k \text{ Candidate Subsequences} \]

\[ \mathcal{L}_1(X, s) \rightarrow \text{Source Sentence} \]

\[ \text{Where can I find that film?} \]

\[ \text{How can I get that picture?} \]

\[ \text{Where can I get that film?} \]

\[ \mathcal{L}_2(X, s) \]

\[ \text{Synonym (similar embeddings)} \]

\[ \arg \max_{X \subseteq V^{(t)}} \mathcal{F}(X) \]

\[ V^{(t)} \rightarrow \hat{X}^* \rightarrow k-\text{sequences} \]

"Where can I get that movie?"

ENCODER

DECODER
DiPS

Induce Diversity while not compromising on Fidelity

Diversity Components

\[ D_1(X) \equiv \text{Rewards unique n-grams} \]


\[ D_2(X) \equiv \text{Rewards Structural Coverage} \]

Fidelity Components

\[ V^{(t)}; 3k \text{ Candidate Subsequences} \]

\[ \mathcal{L}_1(X, s) \text{ Source Sentence} \]

\[ \mathcal{L}_2(X, s) \text{ Synonym (similar embeddings)} \]

\[ \frac{argmax}{X \subset V^{(t)}} F(X) \rightarrow k-\text{sequences} \]

“Where can I get that movie?”

ENCODER

DECODER

\[ \text{‘where’, ‘can’, ‘I’, <eos> } \]

\[ \text{<sos> } \]
DiPS

Induce Diversity while not compromising on Fidelity

Diversity Components

\[ D_1(X) \equiv \text{Rewards unique n-grams} \]

\[ D_2(X) \equiv \text{Rewards Structural Coverage} \]

Fidelity Components

\[ V^{(t)}; 3k \text{ Candidate Subsequences} \]

\[ \mathcal{L}_1(X, s) \rightarrow \mathcal{L}_2(X, s) \]

Source Sentence

Where can I get that movie?

\[ \text{argmax}_{X \in V^{(t)}} \mathcal{F}(X) \]

k-sequences

ENCODER

DECODER
Diversity Components

\[ V^{(t)}: 3k \text{ Candidate Subsequences} \]

Where can I find that film?
How can I get that picture?
Where can I get that film?

\[ D_1(X) \equiv \text{Rewards unique n-grams} \]

Where ---- can I find ---- that picture
------- How can I ---- get that picture

\[ D_2(X) \equiv \text{Rewards Structural Coverage} \]
Diversity Components

\[ V^{(t)}: \text{3k Candidate Subsequences} \]

Where can I find that film?

How can I get that picture?

Where can I get that film?

\[ \beta_n \sum_{n=1}^{N} \bigcup_{x \in X} x_{n-gram} \]

Diversity Components

\[ D_1(X) \equiv \text{Rewards unique n-grams} \]

N-gram uniqueness

\[ \text{where}, \text{can}, \text{film}, \text{I}, \text{How}, \text{find that}, \text{that picture}, \]

\[ 'I get', 'can I', 'Where can I' \]

\[ D_2(X) \equiv \text{Rewards Structural Coverage} \]
Diversity Components

V(t): 3k Candidate Subsequences

Where can I find that film?
How can I get that picture?
Where can I get that film?

D_1(X) \equiv \text{Rewards unique n-grams}

\text{Diversity Components}

\text{N-gram uniqueness}
\sum_{n=1}^{N} \beta^n \bigcup_{x \in X} x_{n-gram}

D_2(X) \equiv \text{Rewards Structural Coverage}

\text{Structural Coverage}
\sum_{x_i \in V(t)} \sum_{x_j \in X} R(x_i, x_j)

R(x_i, x_j) = 1 - \frac{\text{EditDistance}(x_i, x_j)}{|x_i| + |x_j|}
DiPS

Induce Diversity while not compromising on Fidelity

Diversity Components

\[ D_1(X) \equiv \text{Rewards unique n-grams} \]


Fidelity Components

\[ V^{(t)}: 3k \text{ Candidate Subsequences} \]

\[ D_2(X) \equiv \text{Rewards Structural Coverage} \]

\[ \text{Synonym (similar embeddings)} \]

\[ \mathcal{L}_1(X, s) \]

\[ \text{Source Sentence} \]

\[ \text{Where can I get that movie?} \]

\[ \text{Where can I get that film?} \]

\[ \text{Where can I find that picture?} \]

\[ \text{How can I get that picture?} \]

\[ \text{argmax}_{X \subseteq V^{(t)}} \mathcal{F}(X) \]

\[ \mathcal{L}_2(X, s) \]

\[ \text{ENCODER} \]

\[ \text{DECODER} \]

\[ \rightarrow k-\text{sequences} \]
DiPS
Induce Diversity while not compromising on Fidelity

$\mathcal{D}_1(X) \equiv$ Rewards unique n-grams

$\mathcal{D}_2(X) \equiv$ Rewards Structural Coverage

$V^{(t)}$: 3k Candidate Subsequences

$\mathcal{L}_1(X, s)$

Source Sentence

“Where can I get that movie?”

$\text{ENCODER}$

$\text{DECODER}$

$k$-sequences
Fidelity Components

$V^{(t)}$: 3k Candidate Subsequences

Source Sentence

$L_1(X, s)$

$L_2(X, s)$

Synonym (similar embeddings)

Where can I find that film?

How can I get that picture?

Where can I get that film?
Fidelity Components

V^{(t)}: 3k Candidate Subsequences
Where can I find that film?
How can I get that picture?
Where can I get that film?

L_1(X, s) \rightarrow \text{Source Sentence} \rightarrow L_2(X, s)

Fidelity Components

Lexical Similarity

\sqrt{\sum_{x\in X} \sum_{n=1}^{N} \beta^n |x_n-gram \cap s_n-gram|}
Fidelity Components

V(t): 3k Candidate Subsequences

Where can I find that film?

How can I get that picture?

Where can I get that film?

L1(X, s) → Source Sentence → L2(X, s)

Synonym (similar embeddings)

Lexical Similarity

\[ \sum_{x \in X} \sum_{n=1}^{N} \beta^n |x_{n-gram} \cap s_{n-gram}| \]

Embedding based Similarity

\[ \sqrt{\sum_{x \in X} \mathcal{S}(x, s)} \]

\[ \mathcal{S}(x, s) = \frac{1}{|x|} \sum_{w_j \in x} \max_{w_i \in s} \psi(v_{w_i}, v_{w_j}) \]
DiPS Objective

Diversity Components

\[ D_1(X) \equiv \text{Rewards unique n-grams} \]

\['where', 'can', 'film', 'I', 'How', 'find that', 'that picture', 'I get', 'can I', 'Where can I'\]

Fidelity Components

\[ V^{(t)}: 3k \text{ Candidate Subsequences} \]

\[ L_1(X, s) \]

\[ L_2(X, s) \]

Source Sentence

Where can I get that movie?

\[ \text{Rewards Structural Coverage} \]

\[ \text{Synonym (similar embeddings)} \]
DiPS Objective

$$\arg\max_{X \subseteq V, X \in \langle k \rangle} F(X)$$

$$F(X) = \lambda(\mu_1 D_1(X) + \mu_2 D_2(X)) + (1 - \lambda)(\nu_1 L_1(X, s) + \nu_2 L_2(X, s))$$

**Diversity Components**


$$D_1(X) \equiv \text{Rewards unique n-grams}$$

**Fidelity Components**

- $V^{(t)}$: 3k Candidate Subsequences
- $L_1(X, s)$
- $L_2(X, s)$

- Source Sentence: Where can I get that movie?
- Synonym (similar embeddings)

**Rewards Unique n-grams**

Where ······ can I find ······ that picture

- How can I ······ get that picture

$$D_2(X) \equiv \text{Rewards Structural Coverage}$$
DiPS Objective

$$\text{argmax}_{X \subseteq V, X = |k|} F(X)$$

$$F(X) = \lambda (\mu_1 D_1(X) + \mu_2 D_2(X)) + (1 - \lambda) (\nu_1 L_1(X, s) + \nu_2 L_2(X, s))$$

**Diversity Components**


**Fidelity Components**

**$V^{(t)}$: 3k Candidate Subsequences**

- Where can I find that film?
- How can I get that picture?
- Where can I get that film?

**$\mathcal{D}_1(X)$** \(\equiv\) Rewards unique n-grams

**$\mathcal{D}_2(X)$** \(\equiv\) Rewards Structural Coverage

**$\mathcal{L}_1(X, s)$**

**Source Sentence**

- Where can I get that movie?

**$\mathcal{L}_2(X, s)$**

**Synonym (similar embeddings)**
DiPS Objective

\[
\text{argmax}_{X \subseteq V, |X|=k} F(X)
\]

\[
F(X) = \lambda (\mu_1 D_1(X) + \mu_2 D_2(X)) + (1 - \lambda) (\nu_1 L_1(X, s) + \nu_2 L_2(X, s))
\]

Diversity Components

\[
D_1(X) \equiv \text{Rewards unique n-grams}
\]

\[
D_2(X) \equiv \text{Rewards Structural Coverage}
\]

Fidelity Components

\[
V^{(t)}; 3k \text{ Candidate Subsequences}
\]

Source Sentence

\[
\mathcal{L}_1(X, s) \quad \mathcal{L}_2(X, s)
\]

\[
\text{Where can I find that film?}
\]

\[
\text{How can I get that picture?}
\]

\[
\text{Where can I get that film?}
\]

\[
\text{Synonym (similar embeddings)}
\]
Fidelity & Diversity
(Quora Dataset)
Fidelity & Diversity
(Quora Dataset)

BLEU (Fidelity)

SBS
DBS
VAE-SVG
DPP
SSR
DiPS (Ours)
Fidelity & Diversity
(Quora Dataset)

BLEU (Fidelity)

Models

4-Distinct (Diversity)
Fidelity & Diversity
(Quora Dataset)

DiPS induces diversity without compromising on fidelity
Data Augmentation
Paraphrase Detection

Quora Dataset

<table>
<thead>
<tr>
<th>Models</th>
<th>No Aug</th>
<th>SBS</th>
<th>DPP</th>
<th>SSR</th>
<th>DBS</th>
<th>DiPS (Ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>67</td>
<td>70</td>
<td>73</td>
<td>76</td>
<td>73</td>
<td>76</td>
</tr>
</tbody>
</table>

Bar chart showing accuracy comparison between different models with and without data augmentation.
DiPS data augmentation helps in paraphrase detection.
Data Augmentation for Intent Classification

Dataset: SNIPS

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>93</td>
<td>94</td>
<td>95</td>
<td>96</td>
<td>97</td>
<td>97</td>
</tr>
</tbody>
</table>

- **No. Aug**: No data augmentation
- **SBS**: Simple data augmentation
- **DBS**: Deep data augmentation
- **Syn. Rep**: Synthetic representation
- **Cont. Aug**: Content augmentation
- **DiPS (Ours)**: DiPS (Our proposed solution)
Data Augmentation for Intent Classification

Dataset: SNIPS

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LogReg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data augmentation using DiPS improves intent classification
Data Augmentation for Intent Classification

Dataset: Yahoo-L31

Models:
- No Aug.
- SBS
- DBS
- Cont. Aug.
- DiPS (Ours)

Accuracy:
- LogReg
- LSTM

Graph showing the accuracy comparison between different data augmentation methods and models.
Data Augmentation for Intent Classification

Data augmentation using DiPS improves intent classification
Conclusion
Conclusion

Problem

Diversity in Paraphrases

Without compromising on fidelity
Conclusion

**Problem**

Diversity in Paraphrases

Without compromising on fidelity

**Method**

DiPS

Sub-modular optimisation
Conclusion

Problem
Diversity in Paraphrases
Without compromising on fidelity

Method
DiPS
Sub-modular optimisation

Take-Aways
Seq2Seq + Diversity
Data Augmentation Using Paraphrasing
Code

https://github.com/malllabiisc/DiPS
Code

https://github.com/malllabiisc/DiPS

Acknowledgement
Code
https://github.com/malllabiisc/DiPS

Acknowledgement

ashutosh@iisc.ac.in, satwik55@gmail.com
Code
https://github.com/malllabiisc/DiPS

Acknowledgement

ashutosh@iisc.ac.in, satwik55@gmail.com

Thank you