LREC-COLING 2024

Controllable Paraphrase Generation for Semantic and Lexical Similarities

Yuya Ogasa[1], Tomoyuki Kajiwara[2], Yuki Arase[1]

[1]Osaka University, [2]Ehime University

Paraphrase Generation

- Paraphrases render the meaning of text using different words, phrases, and syntactic structures.

Original Sentences	Paraphrases
The news made him happy.	He was happy to hear the news.
I choose a big box from several alternatives.	I select a large box from several alternatives.

- Paraphrase generation contributes to **data augmentation**.

  ex. machine translation (Effendi et al., 2018)

  task-oriented dialog systems (Jolly et al., 2020)

Lexically Diverse Paraphrase

Crucial for data augmentation

- They enhance the linguistic diversity of the original corpus (Qian et al., 2019).

Examples of Paraphrases by Lexical Diversity

Original Sentences	Paraphrases	Lexical Diversity
The news made him happy.	He was happy to hear the news.	O
I choose a big box from several alternatives.	I select a large box from several alternatives.	×

Lexically Diverse Paraphrase

Crucial for data augmentation

- They enhance the linguistic diversity of the original corpus (Qian et al., 2019).

Examples of Paraphrases by Lexical Diversity

Original Sentences	Paraphrases	Lexical Diversity
The news made him happy.	He was happy to hear the news.	O
I choose a big box from several alternatives.	I select a large box from several alternatives.	×

Many words change yet the meaning is intact.

Lexically Diverse Paraphrase

Crucial for data augmentation

- They enhance the linguistic diversity of the original corpus (Qian et al., 2019).

Examples of Paraphrases by Lexical Diversity

Original Sentences	Paraphrases	Lexical Diversity
The news made him happy.	He was happy to hear the news.	○
I choose a big box from several alternatives.	I select a large box from several alternatives.	×

Only a few words change.

Generating Lexically Diverse Paraphrases

Surface changes easily make sentences semantically less similar.
(Bandel et al., 2022)

Generating Lexically diverse paraphrases is challenging.

Original Sentences	Paraphrases	Lexical Diversity
The news made him happy.	He was happy to hear the news.	○
I **choose** a **big** box from several alternatives.	I **select** a **large** box from several alternatives.	×

Define Lexically Diverse Paraphrase in Our Study

- Lexically diverse paraphrase pairs have **high Semantic and low Lexical Similarities.**
  - Semantic Similarity : The estimated score of the fine-tuned pre-trained model
  - Lexical Similarity : sentence BLEU score

Examples of Paraphrases by Lexical diversity

Original Sentences	Paraphrases	Lexical Diversity	Semantic Similarity	Lexical Similarity
The news made him happy.	He was happy to hear the news.	○	High	Low
I choose a big box from several alternatives.	I select a large box from several alternatives.	×	High	High

Existing Paraphrases vs Our Paraphrases

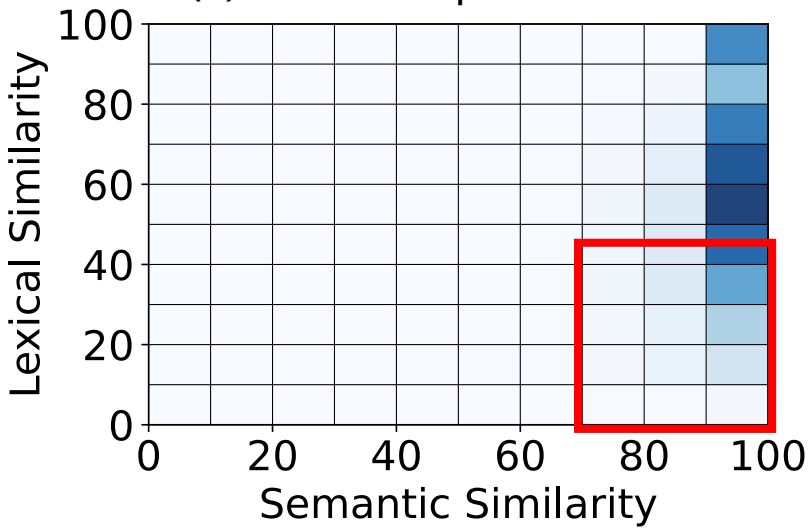
Heatmaps illustrate the distributions of semantic and lexical similarities among paraphrases.

Darker cell colors indicate higher ratios of paraphrases.

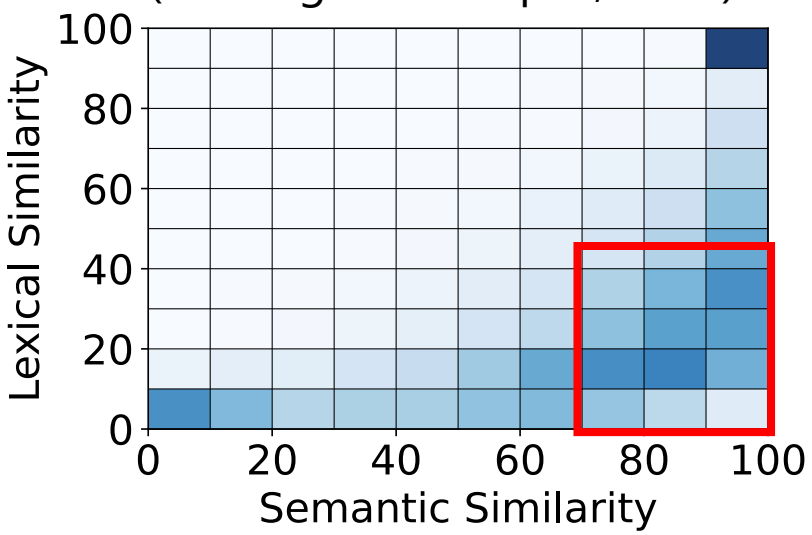
(a),(b) : Existing Paraphrases, (c) : Our Paraphrases

: Lexically diverse paraphrase

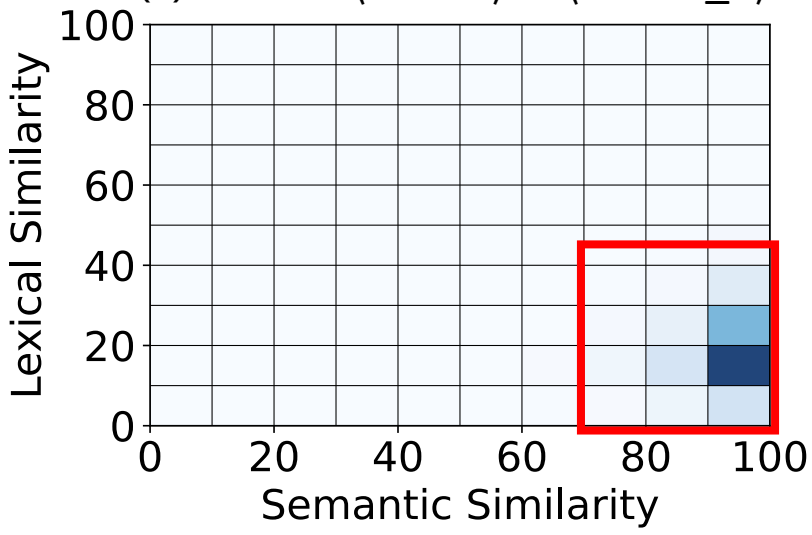
(a) Round-trip translation



(b) ParaNMT- 50M
(Wieting and Gimpel., 2018)



(c) Ours : ⟨SIM95⟩ & ⟨BLEU0_5⟩



Existing Paraphrases vs Our Paraphrases

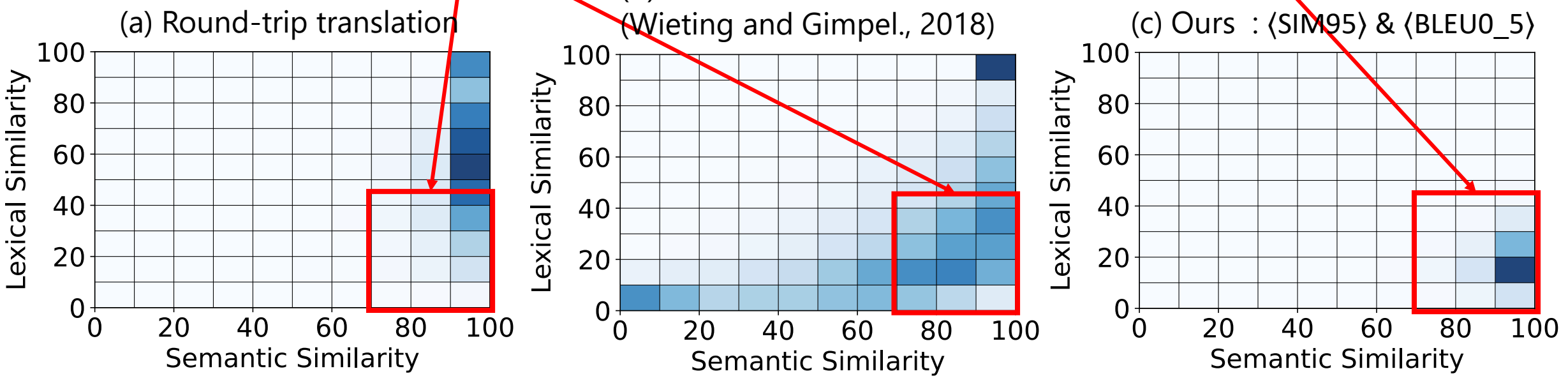
Darker cell colors indicate higher ratios of paraphrases.

(a),(b) : Existing Paraphrases, (c) : Our Paraphrases

: Lexically diverse paraphrase

Fewer lexically diverse paraphrases

Many lexically diverse paraphrases

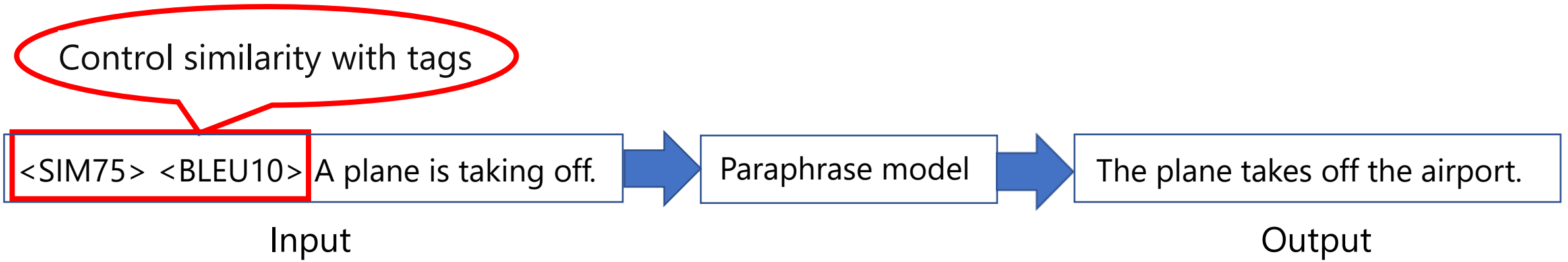


Suitable Paraphrasing for Data Augmentation

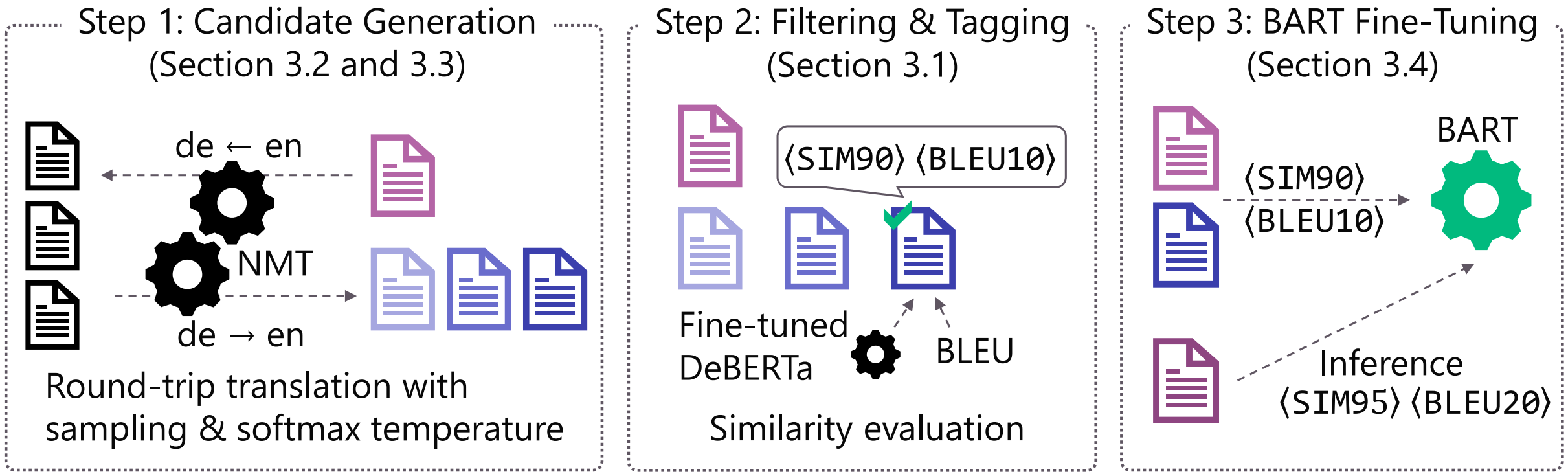
- Data augmentation effects vary with semantic and lexical similarity, influenced by the task (shown by our experiments).

- Problem
  - Existing paraphrase methods lacked easy similarity control.

Our method

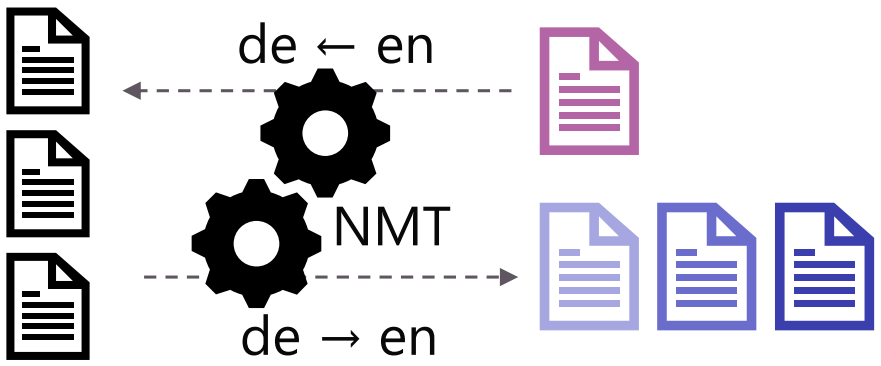


Proposed Method



Proposed Method ~ Controllable Generation ~

Step 1: Candidate Generation
(Section 3.2 and 3.3)

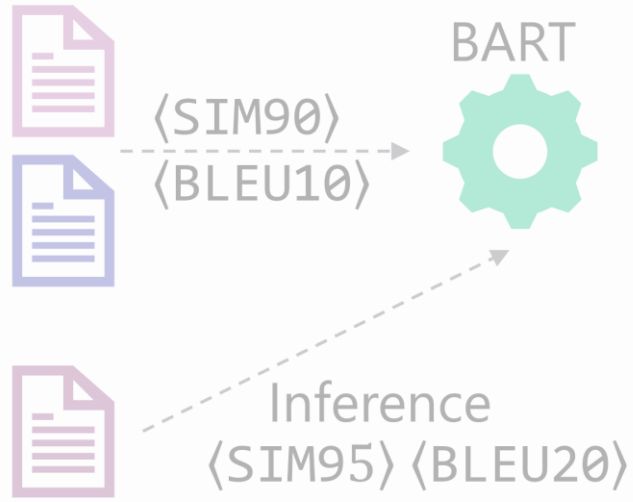


Round-trip translation with
sampling & softmax temperature

Step 2: Filtering & Tagging
(Section 3.1)

Similarity evaluation

Step 3: BART Fine-Tuning
(Section 3.4)



Proposed Method ~Candidate Generation~

Round-trip translation

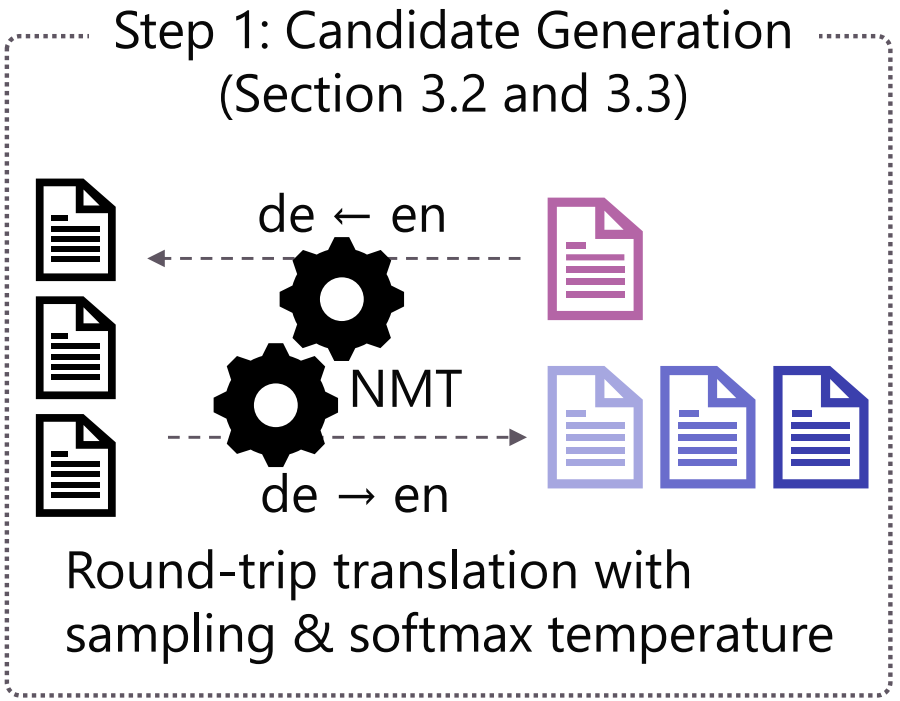
- Lexical similarity of the paraphrased sentence pairs: high

Solution: Round-trip translation **with sampling & softmax temperature**

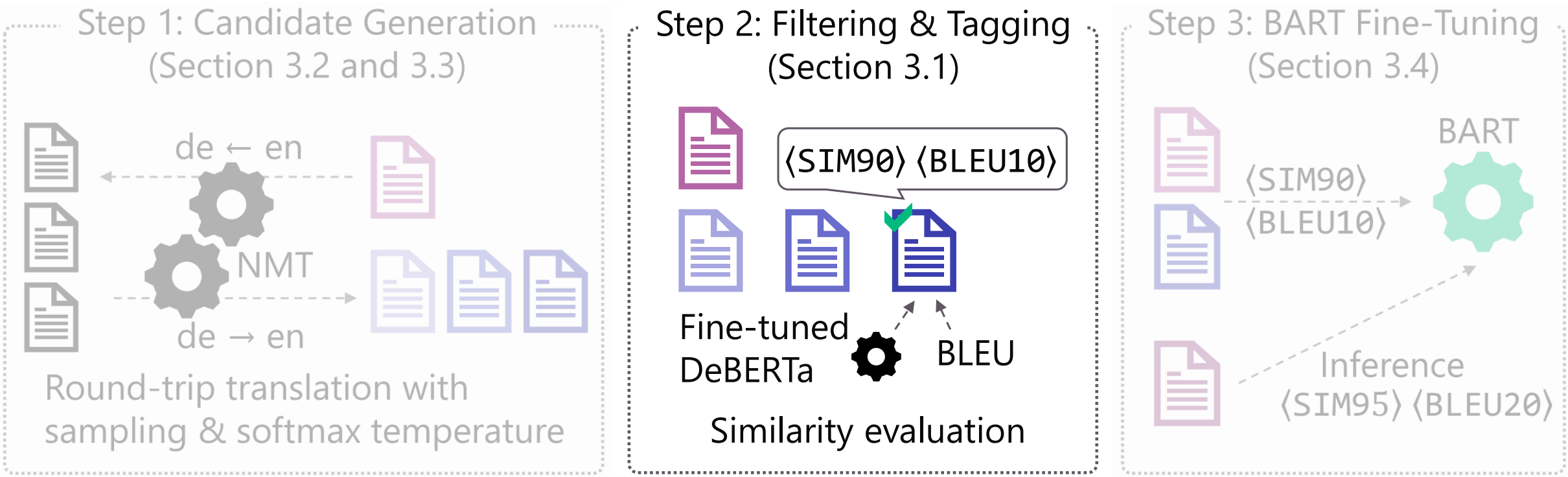
- Tendency : Lexical Similarity ↓ but Semantic Similarity ↓

Using the method, generating 120M candidates

Extracting 5M lexically diverse paraphrases (Step 2)



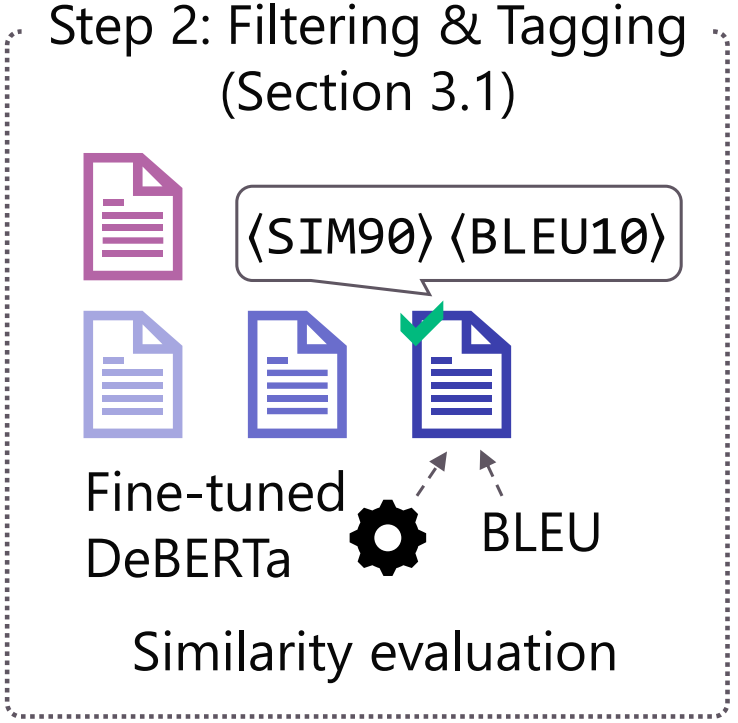
Proposed Method ~ Controllable Generation ~



Proposed Method ~Similarity Metrics~

Semantic Similarity

- Estimated with fine-tuned pre-trained models.
  - Model : DeBERTaV3
  - Training data : STS-B



Lexical Similarity

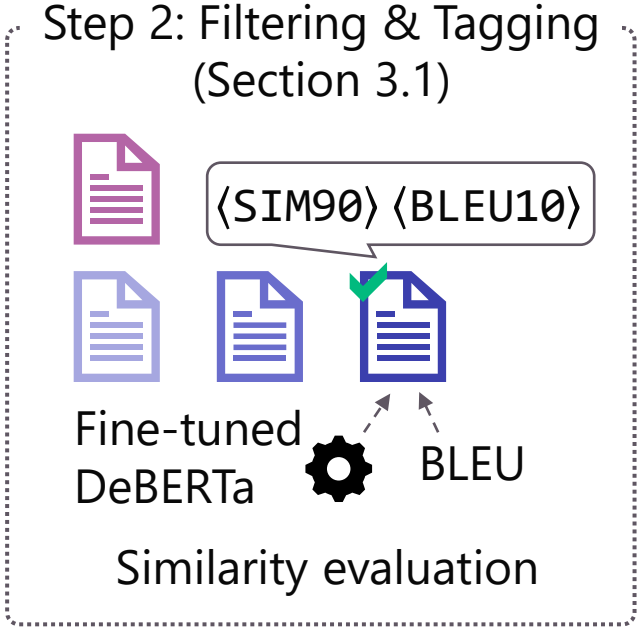
- Sentence BLEU
  - This metric has been commonly used to assess lexical similarity (diversity).
    (David Chen and William Dolan, 2011) (Tian et al.,2017) (Jiang et al., 2020)

Proposed Method ~Filtering & Tagging~

- Both Semantic and Lexical similarity scores have a range of [0, 100].
  - The higher the number, the more similar.

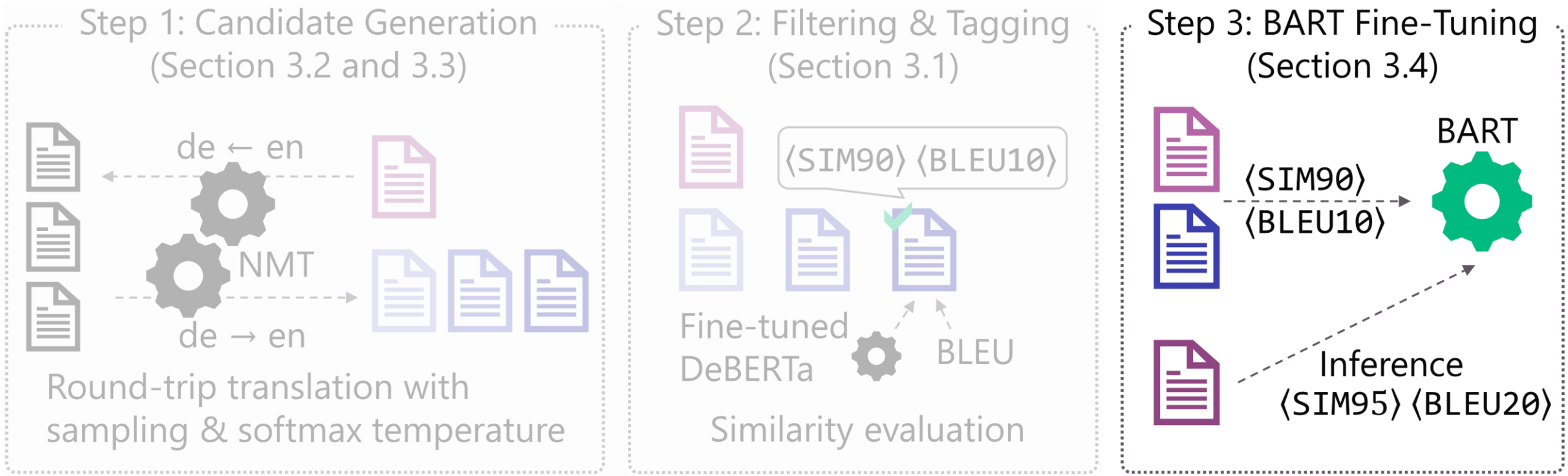
- Our definition of lexically diverse paraphrases
  - Semantic similarity > 70
  - Lexical similarity ≤ 45

- Semantic and lexical similarities tagged in intervals of 5
  - Semantic: ⟨SIM70⟩, ⟨SIM75⟩, . . . , ⟨SIM95⟩ (6 tags)
  - Lexical: ⟨BLEU0_5⟩, ⟨BLEU10⟩, ⟨BLEU15⟩, . . . , ⟨BLEU40⟩ (8 tags)



Following these metrics, we extract lexically diverse paraphrases from step1 candidates.

Proposed Method ~ Controllable Generation ~

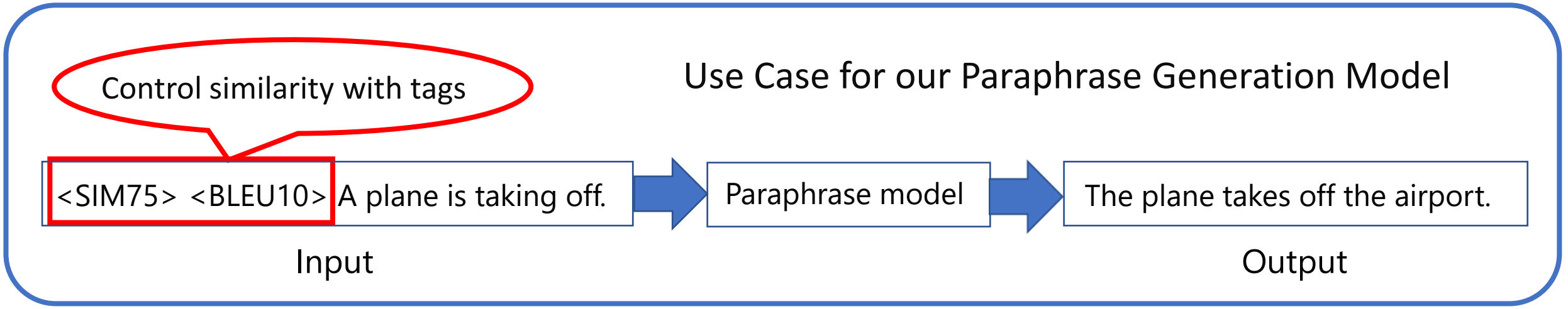


Proposed Method ~BART Fine-Tuning~

Input : Semantic Similarity tag + Lexical Similarity tag + Original Sentence

Output : Paraphrases with Controlled Similarity

Constructed a lexically diverse paraphrase generation model by fine-tuning BART.



Example of Our Paraphrases

- ⟨SIM95⟩ generates almost equivalent meanings to the original.
- ⟨SIM70⟩ introduces moderate changes, like 'leg' to 'ankle' and adding 'the tournament' or 'the French open'.

Original Sentence	Maria Sharapova has been forced to withdraw with a leg injury .
Tags	**Paraphrases**
⟨SIM95⟩ ⟨BLEU0_5⟩	Maria Sharapova **withdrew** with an injury to **her** leg.
⟨SIM95⟩ ⟨BLEU40⟩	Maria Sharapova **had to pull out** with a leg injury.
⟨SIM70⟩ ⟨BLEU0_5⟩	Maria Sharapova **withdrew from the tournament** with an **ankle** injury.
⟨SIM70⟩ ⟨BLEU40⟩	Maria Sharapova has been forced to **pull out of the French Open** with a leg injury.

Validating the Effectiveness of Our Paraphrase-Driven Data Augmentation.

1. Contrastive Learning : SimCSE (Gao et al., 2021)
2. Pre-fine-tuning : STILTs (Phang et al., 2018)

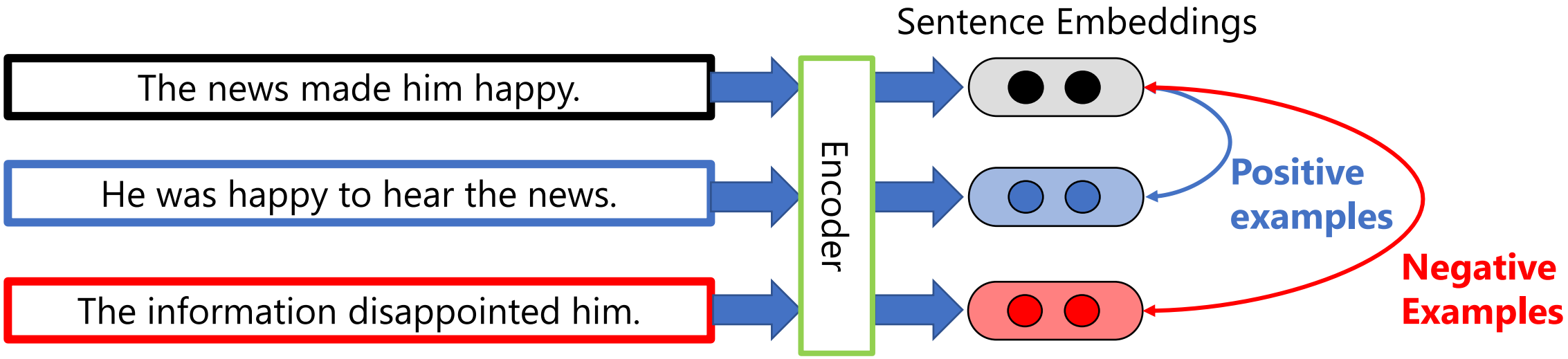
Comparative Methods

- Round-trip translation (RTT)

SimCSE

- **Sim**ple **C**ontrastive Learning of **S**entence **E**mbeddings (Gao et al., 2021)

- Positive embedding for similar meanings, negative for different.
- Fine-tune pre-trained models to attract positives and separate negatives.



Doubled training data using our method.

- Training data : natural language inference (NLI) dataset
- Task: unsupervised 7 STS tasks
- Metric: Spearman's rank correlation coefficient (ρ)
- Evaluation model: BERT-base
- Tags used in our model
  - Semantic Similarity: ⟨SIM95⟩ (Fixing at high semantic similarity)
  - Lexical Similarity: ⟨BLEU0_5⟩, ⟨BLEU20⟩, ⟨BLEU40⟩
  - Experiment with 3 different combinations

SimCSE~Result~

- Proposed method outperforms comparative method.
- ⟨BLEU0_5⟩ (Ours) improve SimCSE scores by 0.60, RTT scores by 0.39.

Methods	Scores
SimCSE	81.70
RTT	81.91
⟨BLEU0_5⟩ (Ours)	**82.30**
⟨BLEU20⟩ (Ours)	82.18
⟨BLEU40⟩ (Ours)	82.07

STILTs

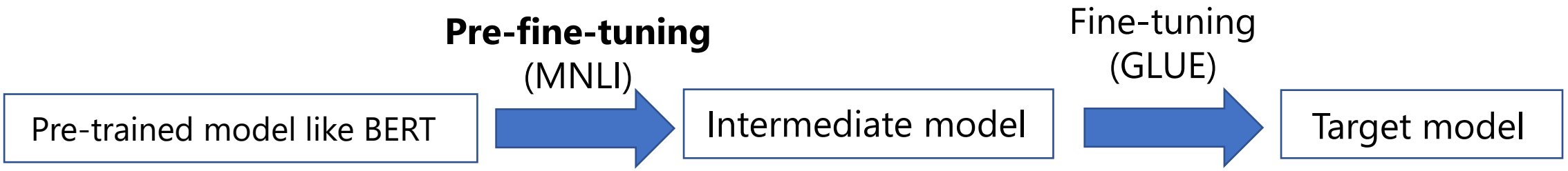
Supplementary **T**raining on **I**ntermediate **L**abeled-data **T**ask**s** (Phang et al., 2018)

- One of pre-fine-tuning method
- Pre-fine-tuning improves the performance of the pretrained language model on downstream tasks.

STILTs ~Setting~

Doubled Pre-fine-tuning data using our method.

- Pre-fine-tuning data: MNLI dataset
- Task: GLUE benchmark (Wang et al., 2018)
- Evaluation model: BERT-large
- Tags used in our model
  - Semantic Similarity: ⟨SIM70⟩, ⟨SIM80⟩, ⟨SIM95⟩
  - Lexical Similarity: ⟨BLEU0_5⟩, ⟨BLEU20⟩, ⟨BLEU40⟩
  - Experiment with 9 different combinations



STILTs ~Result①~

- Proposed method outperforms STILTs in 5 tasks and RTT in 8 tasks.
- **Best performing tag varies by task.**

Controlling Semantic and Lexical Similarities is Crucial.

First half of 5 tasks

Method	CoLA	SST-2	MRPC	QQP	STS-B
BERT	**58.5**	94.3	88.3	**72.4**	86.8
STILTs	57.0	94.2	89.0	71.7	**88.9**
RTT	56.4	94.2	88.6	71.6	88.2
Ours	55.7	**94.9**	**89.2**	71.7	88.7
	⟨SIM70⟩ ⟨BLEU20⟩	⟨SIM70⟩ ⟨BLEU40⟩	⟨SIM95⟩ ⟨BLEU0_5⟩	⟨SIM70⟩ ⟨BLEU20⟩	⟨SIM80⟩ ⟨BLEU40⟩

STILTs ~Result②~

- Proposed method outperforms STILTs in 5 tasks and RTT in 8 tasks.
- **Best performing tag varies by task.**

Controlling Semantic and Lexical Similarities is Crucial.

Second half of 4 tasks

Method	MNLI-m	MNLI-mm	QNLI	RTE
BERT	**86.5**	85.6	92.7	69.0
STILTs	---	---	92.5	79.4
RTT	86.3	86.1	92.4	79.6
Ours	**86.5**	**86.2**	**93.0**	**80.0**
	⟨SIM95⟩ ⟨BLEU20⟩	⟨SIM95⟩ ⟨BLEU0_5⟩	⟨SIM95⟩ ⟨BLEU0_5⟩	⟨SIM95⟩ ⟨BLEU0_5⟩

Conclusion

- We develop **a paraphrase generation model with controllability of semantic and lexical similarities**.
- Extensive experiments about data augmentation confirm the effectiveness of our model.
- We release our model and a corpus of 87 million diverse paraphrases. (https://github.com/Ogamon958/ConPGS)

Check out our paper for:

- Experimental detail settings
- In-depth analysis and discussion of experimental results
- Additional our paraphrase examples