スタイル変換のための
折り返し翻訳に基づく事前訓練

梶原 智之　大阪大学データビリティフロンティア機構
三浦 びわ　株式会社 AI Samurai
荒瀬 由紀　大阪大学大学院情報科学研究科

スタイル変換

Formality Transfer

- カジュアル：I **LOOOVVVEEE** this song **SOOO Much!!!!!!**
- フォーマル：I **very much enjoy** this song.

Text Simplification

- 難解：Alfonso Perez ~~Munoz, usually referred to as Alfonso,~~ is a former Spanish **footballer**~~, in the striker position~~.
- 平易：Alfonso Perez is a former Spanish **football player**.

入力文の意味を保持しつつ意味以外の情報**（スタイル）**を制御する

少資源問題

機械翻訳と同じ：パラレルコーパス上でseq2seqモデルを訓練する

機械翻訳と違う：非常に小規模なパラレルコーパスしか利用できない

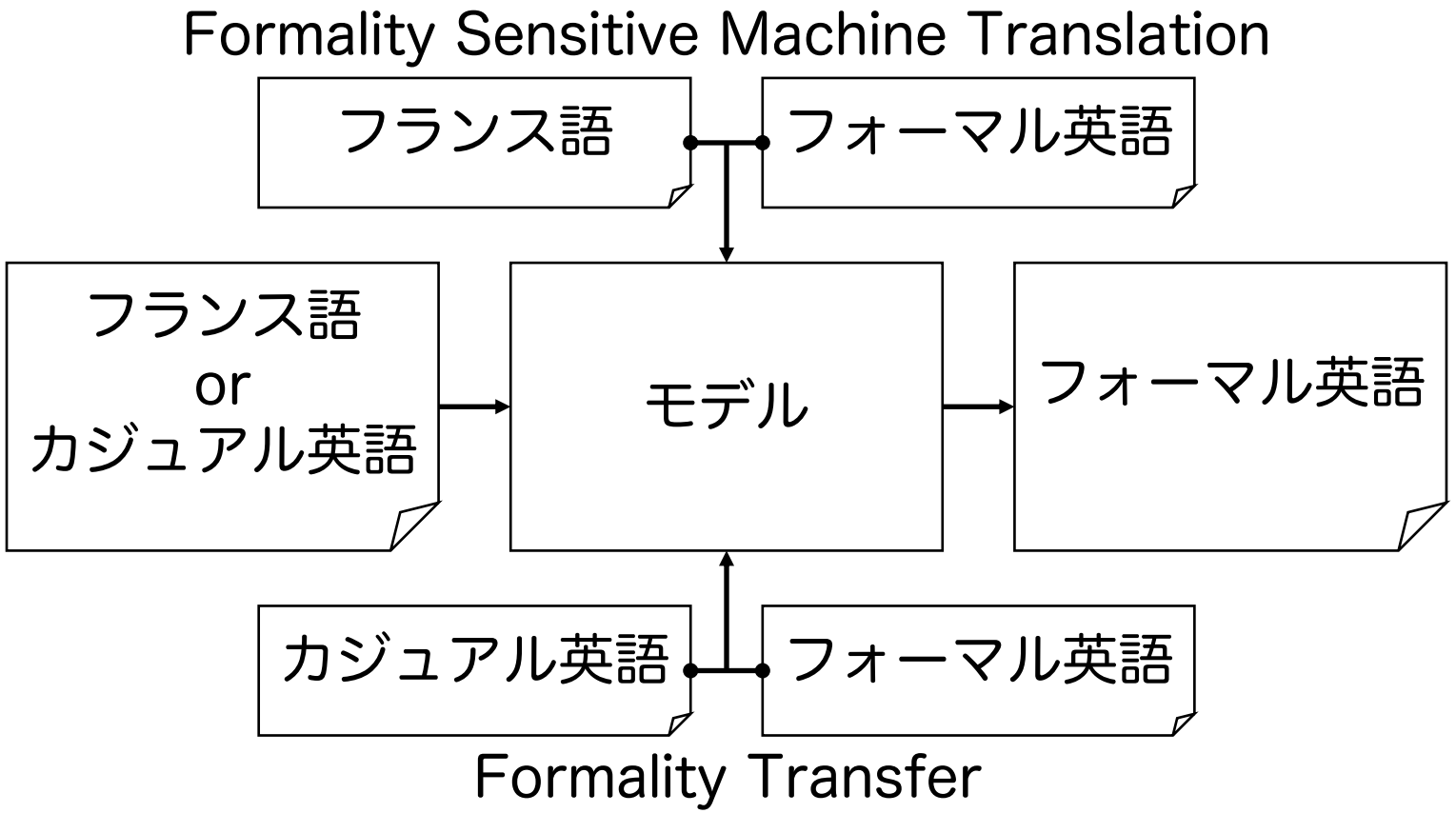
- 対訳データは日々の生活の中で大量に生産／蓄積される
- 単言語パラレルコーパスが自然に作られることは期待できない

機械翻訳	英語 – チェコ語	5,000万文対
	英語 – ドイツ語	500万文対
スタイル変換（英語）	難解 – 平易	10万文対
	カジュアル – フォーマル	10万文対
スタイル変換（日本語）	難解 – 平易	5万文対
	カジュアル – フォーマル	0

先行研究：スタイル変換における少資源問題への対策

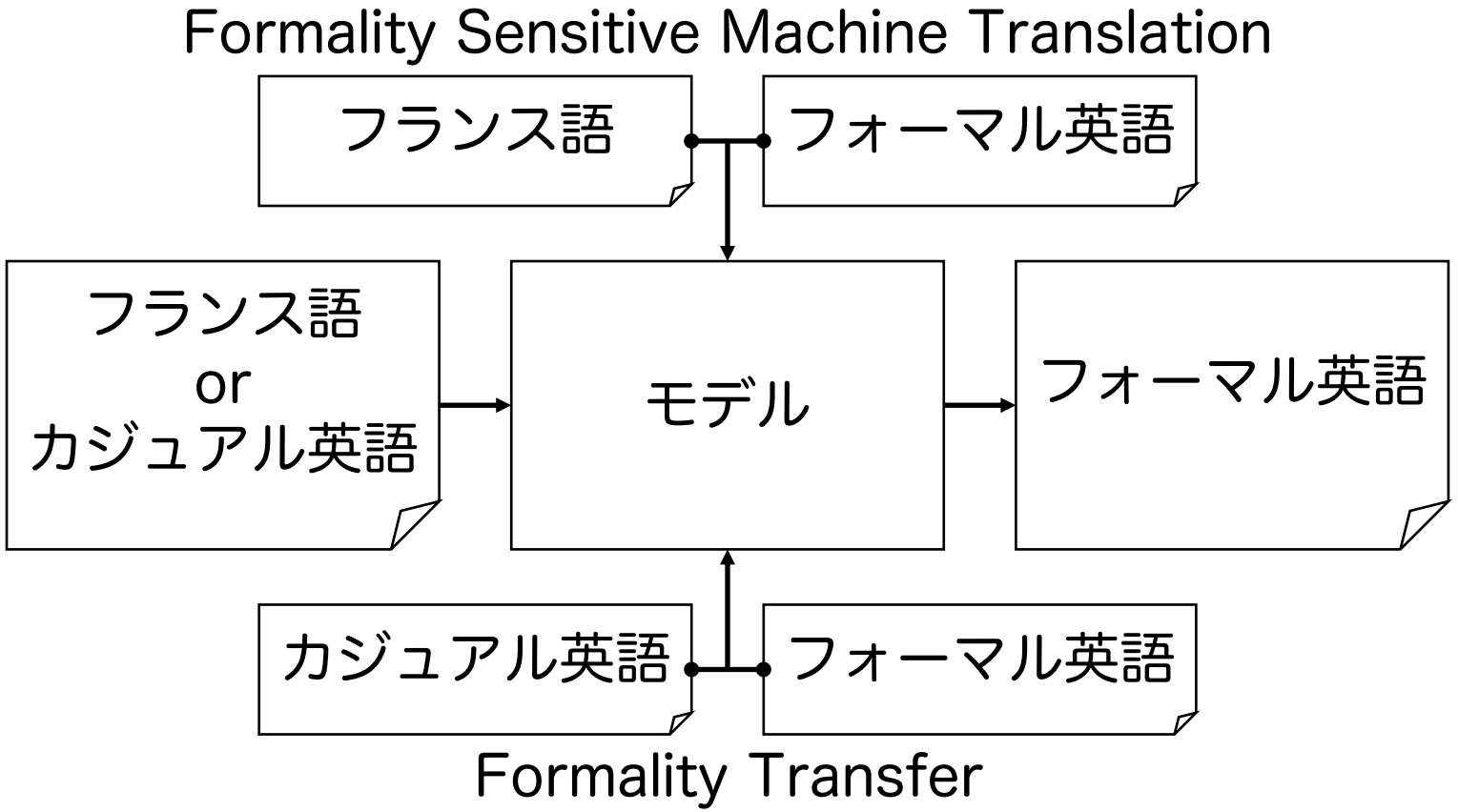
- ルールベースのデータ拡張 [Rao+ 2018]

- 機械翻訳とスタイル変換のマルチタスク学習 [Niu+ 2018]



先行研究：他のスタイルや言語への拡張が困難

- ルールベースのデータ拡張 [Rao+ 2018]
  → スタイルごとに人手でルールを書く必要がある
- 機械翻訳とスタイル変換のマルチタスク学習 [Niu+ 2018]
  → スタイルのラベル付き対訳データを利用できる状況は少ない



本研究：生コーパスを有効活用して少資源問題に対処

スタイル変換における理想的な言い換え

1. 出力文の**スタイル**が適切
2. 出力文が**文法**的に正しい
3. 入出力が**意味**的に等しい

- カジュアル：I **LOOOVVVEEE** this song **SOOO Much!!!!!!**
- フォーマル：I **very much enjoy** this song.

本研究：生コーパスを有効活用して少資源問題に対処

スタイル変換における理想的な言い換え

1. 出力文の**スタイル**が適切
2. 出力文が**文法**的に正しい
3. 入出力が**意味**的に等しい

} **スタイルに依存しない**

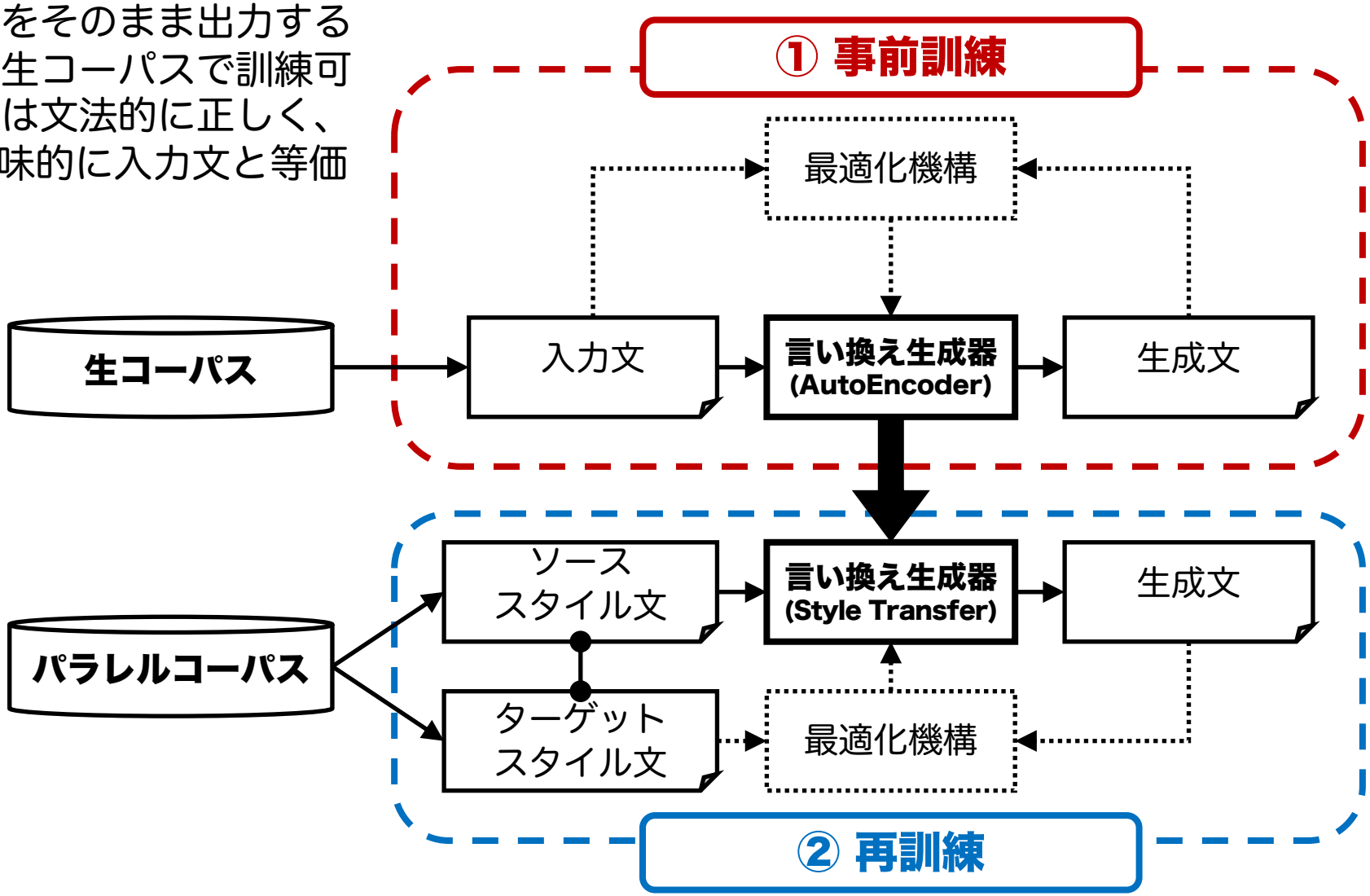
① 事前訓練：**生コーパス**を用いて**意味**と**文法**に関する訓練
- 入力文に対して文法的かつ意味的に等価な文を出力する
- つまり、汎用的な言い換え生成モデルを訓練

② 再訓練：**パラレルコーパス**を用いて**スタイル**に関する訓練
- 入力文に対して目的のスタイルを付与する
- つまり、大量の変換規則からスタイルに適した規則を選出

提案手法１：自己符号化を用いる事前訓練 (AE) + 再訓練

自己符号化: AutoEncoder
- 入力文をそのまま出力する
- 任意の生コーパスで訓練可
- 出力文は文法的に正しく、かつ意味的に入力文と等価

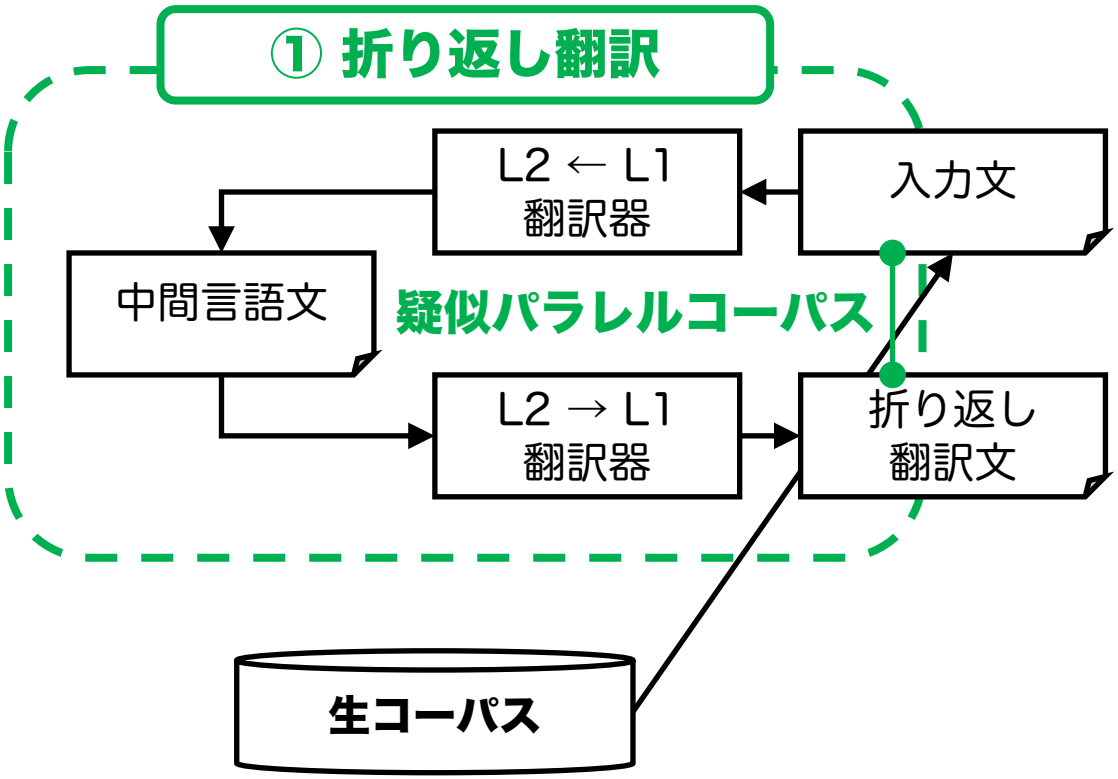
汎用的な言い換え生成モデルを訓練



スタイルに特化した言い換え生成モデルに調整

提案手法２：折り返し翻訳を用いる事前訓練 (RT) + 再訓練

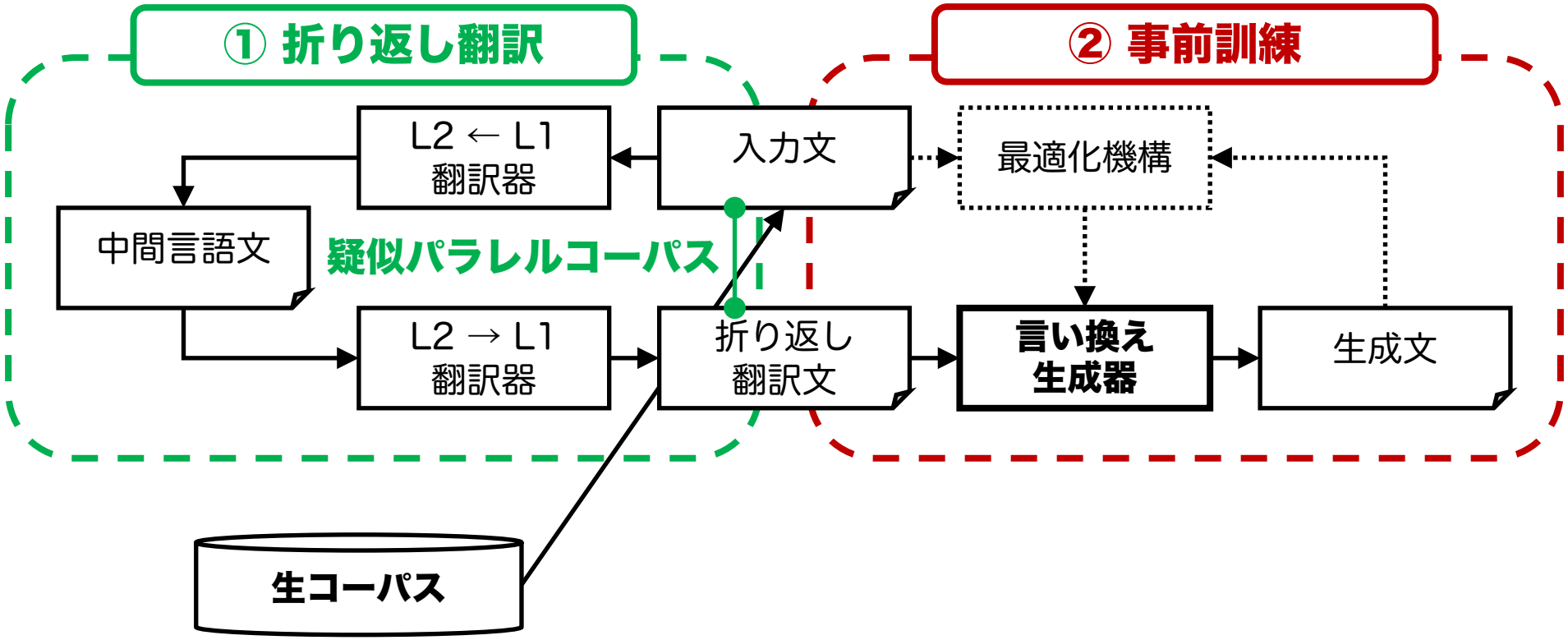
生コーパスから疑似パラレルコーパスを構築



入力文	折り返し翻訳文
I love watching the show.	I love to see the show.
Thanks for asking the question.	Thank you for the question.
The key to a successful relationship is good communication.	Good communication is the key to a successful relationship.

提案手法２：折り返し翻訳を用いる事前訓練 (RT) + 再訓練

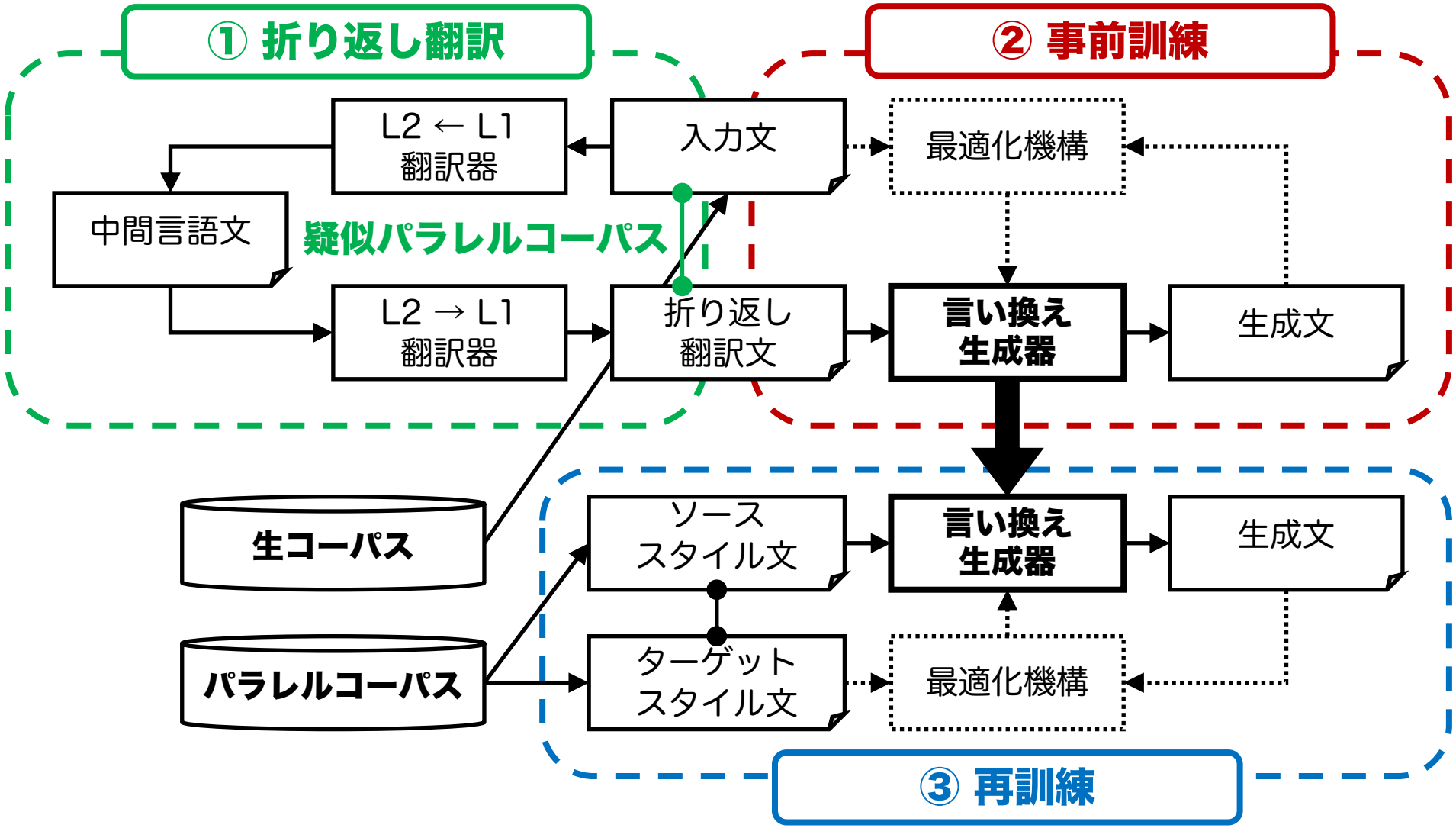
生コーパスから疑似パラレルコーパスを構築　　汎用的な言い換え生成モデルを訓練



入力文	折り返し翻訳文
I love watching the show.	I love to see the show.
Thanks for asking the question.	Thank you for the question.
The key to a successful relationship is good communication.	Good communication is the key to a successful relationship.

提案手法２：折り返し翻訳を用いる事前訓練 (RT) + 再訓練

生コーパスから疑似パラレルコーパスを構築　　汎用的な言い換え生成モデルを訓練



スタイルに特化した言い換え生成モデルに調整

実験設定

スタイル変換

- データ：Yahoo Answersから抽出された**カジュアルな英文**と**フォーマルな英文**のパラレルコーパス (GYAFC)
- モデル：Sockeye上でRNN・CNN・SANの各モデルを構築

		Informal → Formal		Formal → Informal	
	訓練	検証	評価	検証	評価
Entertainment & Music (E&M)	52,595	2,877	1,416	2,356	1,082
Family & Relationships (F&R)	51,967	2,788	1,332	2,247	1,019

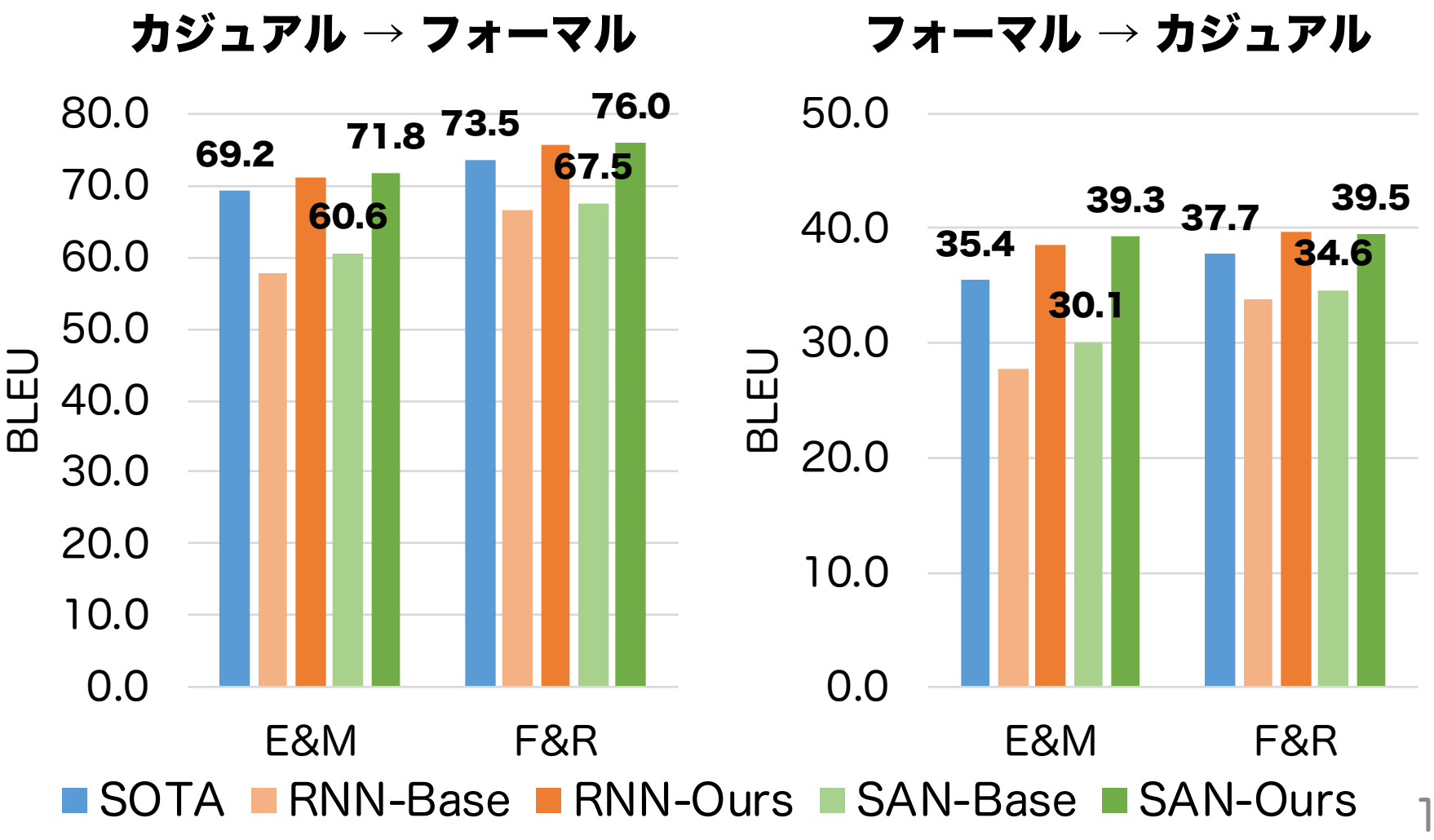
折り返し翻訳

- 生コーパス：Yahoo Answersから300万文
- 翻訳器：スタイル変換器と同じ設定のSANモデル
- データ：WMT2017の英独タスクから450万文対

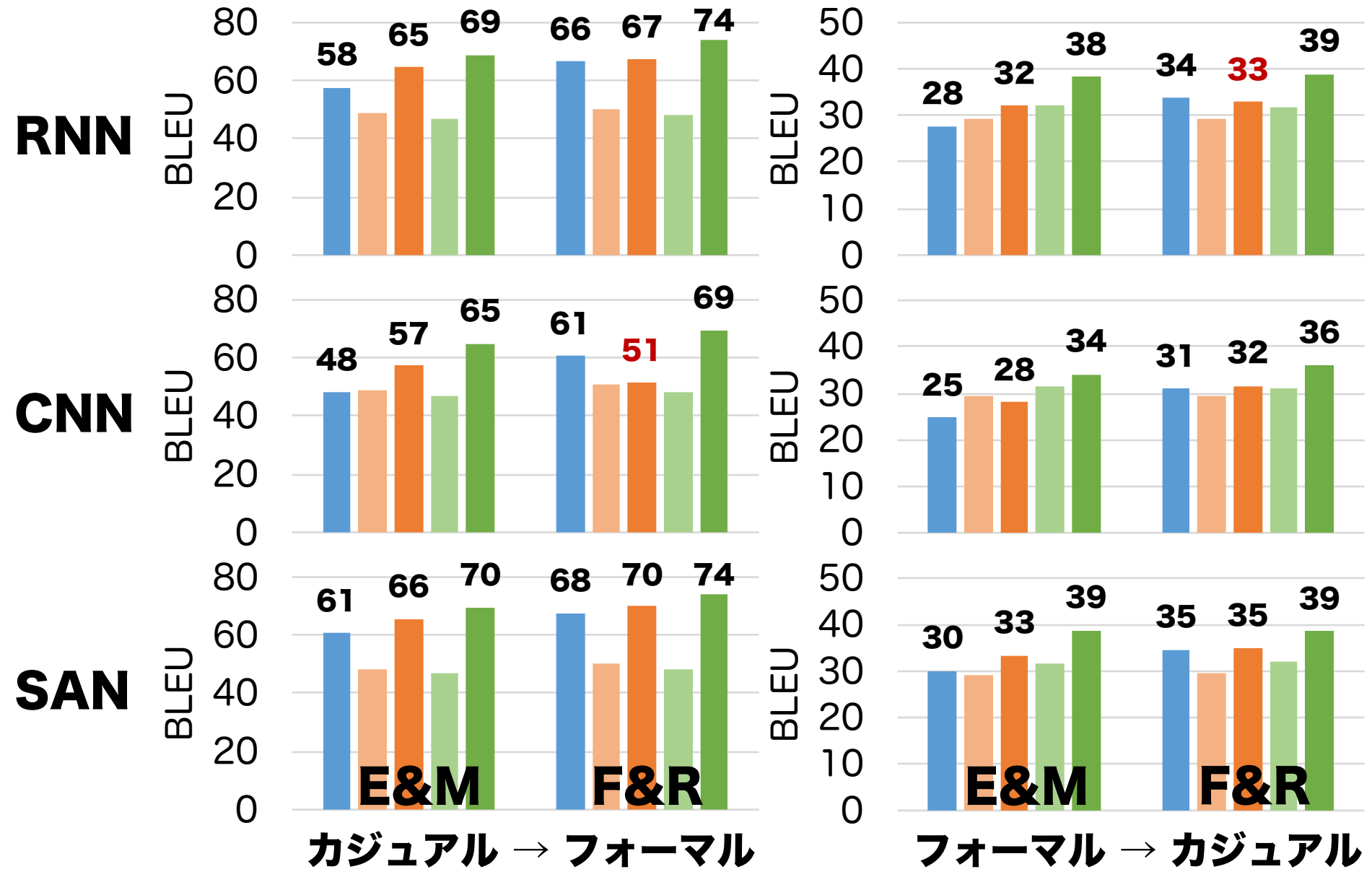
BLEU	英語 → ドイツ語	ドイツ語 → 英語
WMT2017best	26.6	33.5
我々の翻訳器	27.6	33.8

実験結果：折り返し翻訳に基づく提案手法が最高性能を更新

スタイルにもドメインにもベースのモデルにも依存せず
転移学習によって常に大幅に言い換え生成の性能を改善



実験結果：自己符号化に基づく事前訓練も多くの場合に有効



実験結果：スタイル変換の実例

カジュアル → フォーマル

- 入力文　I LOOOOOVVVVVVVEEE this song SOOO Much!!!!!!
- 参照文　I **very much enjoy** this song.
- Ours　I **love** this song **very much**.

先行研究ではカジュアルな表現が残る

- Rao-18　I **loovvvvvvveee** this song so **Much**.
- Niu-18　I **really enjoy VVVVVVVVEEE** this song.

フォーマル → カジュアル

- 入力文　I thoroughly enjoy the hair bands of the 1980s.
- 参照文　I **love** the old hair bands of the **80's!**
- Ours　I **love** the hair bands of the **80’s**.

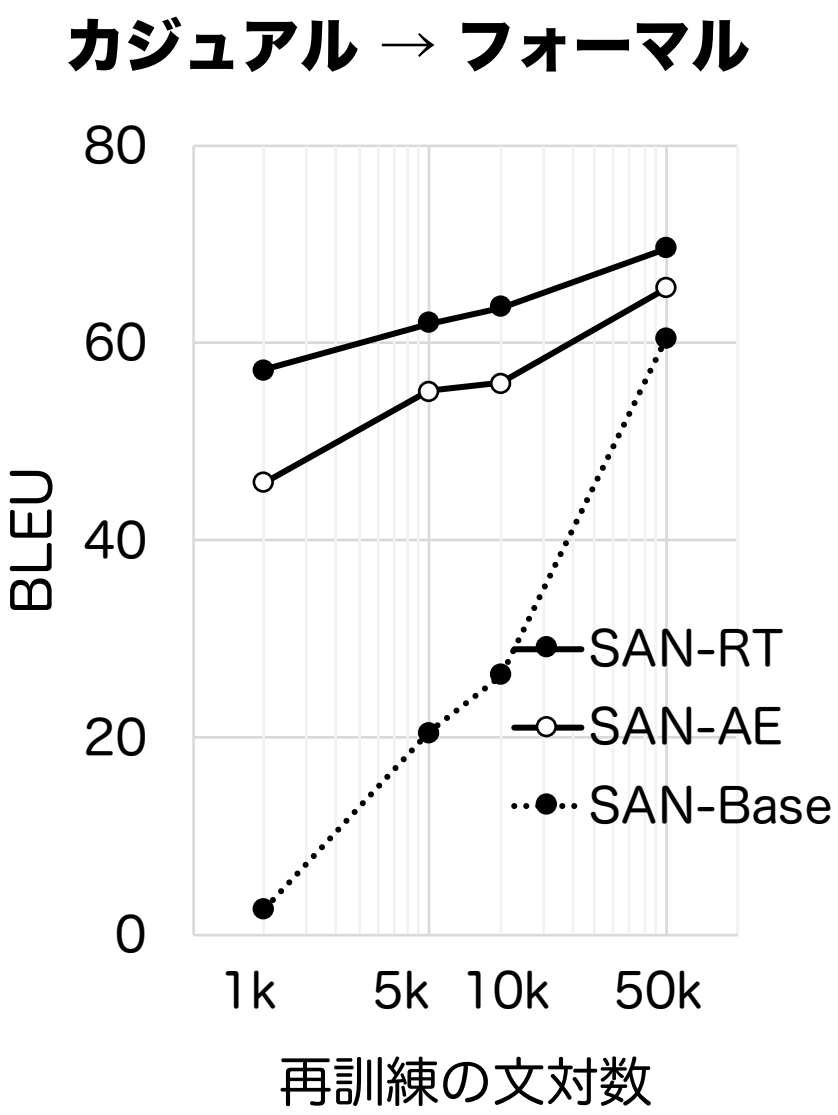
先行研究では意味が変わってしまう

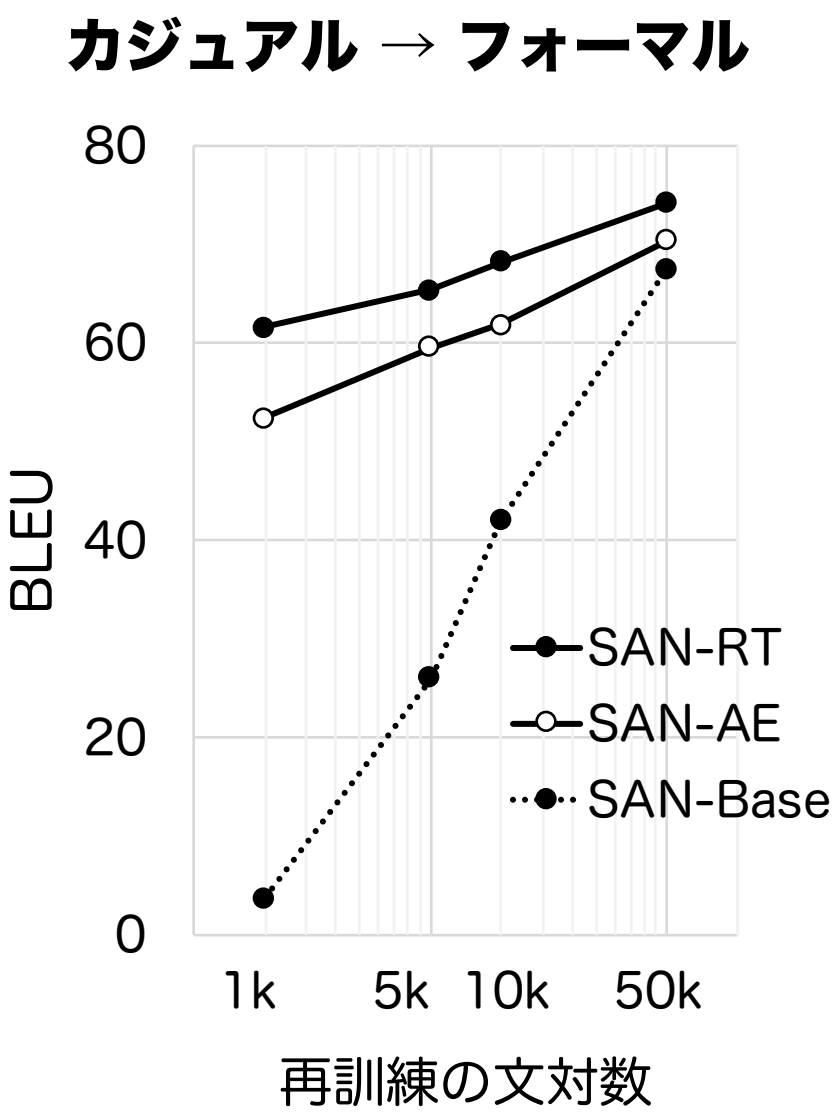
- Rao-18　I just like the **hair of the brids**.
- Niu-18　I **love** the **80’s hair**.

分析と考察

1. 少資源設定の分析
2. 逆翻訳と折り返し翻訳の比較
3. その他のスタイルにおける分析

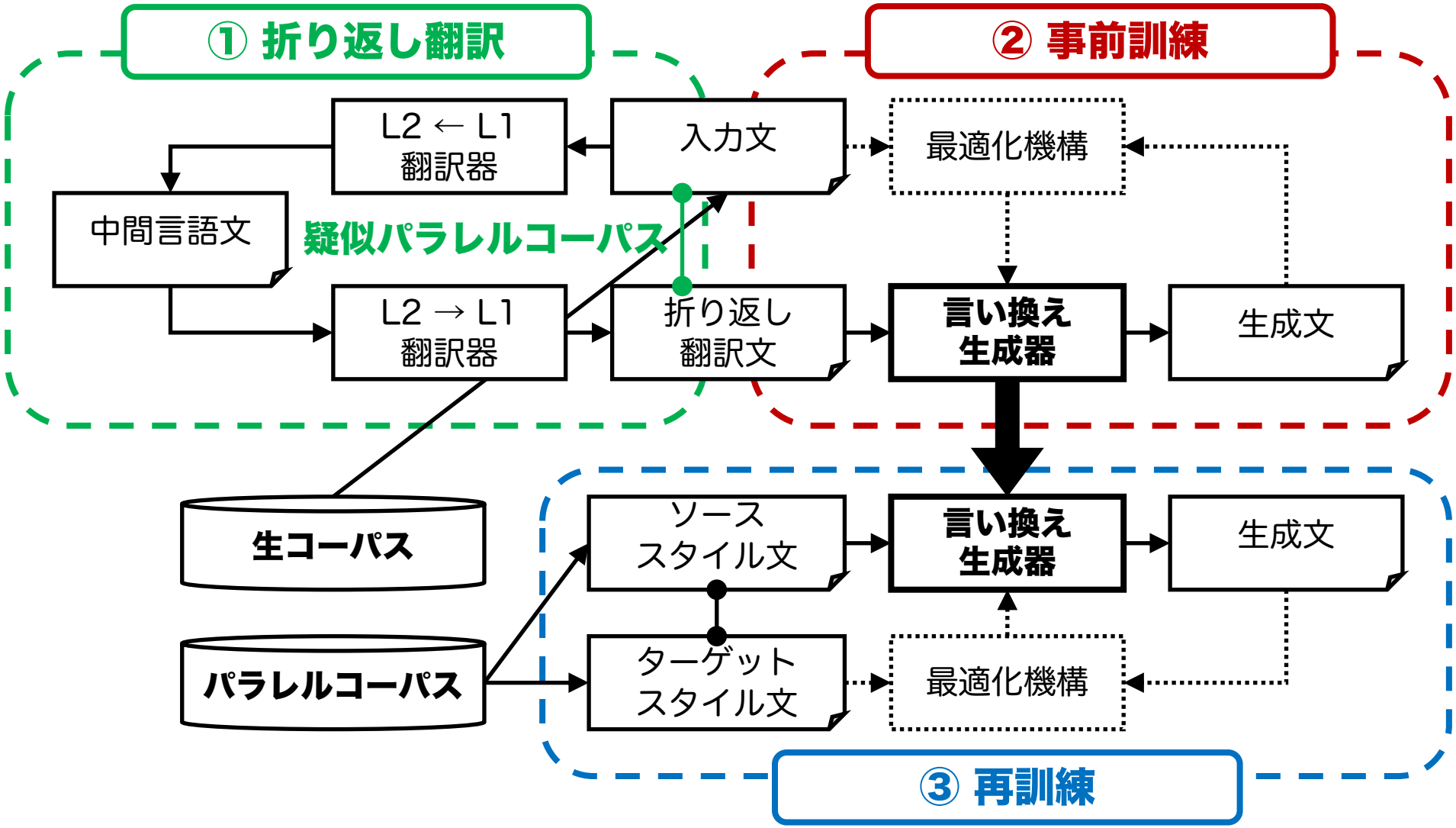
少資源設定の分析：1,000文対でも高品質なスタイル変換を実現





逆翻訳との比較：折り返し翻訳に基づく事前訓練 + 再訓練（再掲）

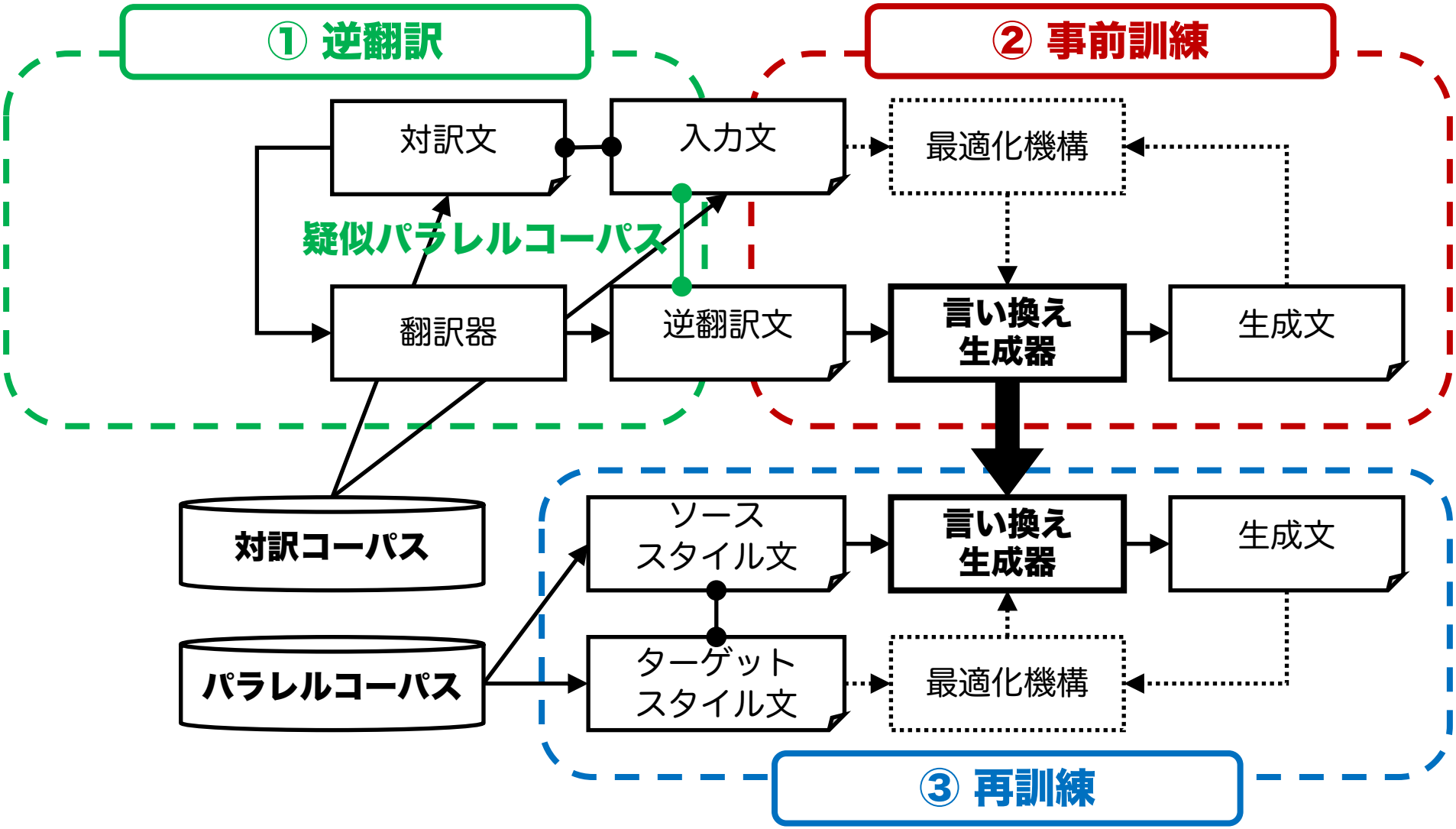
生コーパスから疑似パラレルコーパスを構築　　汎用的な言い換え生成モデルを訓練



スタイルに特化した言い換え生成モデルに調整

逆翻訳との比較：折り返し翻訳を逆翻訳に変更

対訳コーパスから疑似パラレルコーパスを構築　　汎用的な言い換え生成モデルを訓練



スタイルに特化した言い換え生成モデルに調整

逆翻訳との比較：逆翻訳も有効だが、折り返し翻訳の方が良い

カジュアル → フォーマルのBLEU

	E&M	F&R
SAN-Base	60.57	67.52
SAN-AE（自己符号化）	65.57	70.34
SAN-BT（逆翻訳）	69.06	73.39
SAN-RT（折り返し翻訳）	**69.58**	**74.19**

自己符号化 << 逆翻訳

→ 事前訓練において、多様な同義表現を学習することが重要

逆翻訳 < 折り返し翻訳

→ 生コーパスであれば、in-domainのデータを利用できるため

※ 対訳コーパスは任意のドメインにおいて大規模に利用できるわけではない

その他のスタイルにおける分析：丁寧さ以外のスタイルでも有効

テキスト平易化（難解→平易）

- 訓練データ１：WikiSmall（WikipediaとSimple Wikipediaの10万文対）
- 訓練データ２：WikiLarge（WikipediaとSimple Wikipediaの30万文対）
- 検証／評価：Wikipediaを人手で平易化したマルチリファレンス

難解 → 平易のSARI

	WikiSmall	WikiLarge
SAN-Base	34.19	34.58
SAN-RT	**35.46**	**35.99**

先行研究との関係

- スタイル変換におけるデータ拡張 [Rao+ 2018]
- スタイル変換のためのマルチタスク学習 [Niu+ 2018]
  - → **人手や特殊なデータに頼るため他のスタイルへの拡張が困難**
  - ✓ **我々は生コーパスを用いるため他のスタイルへの拡張が容易**
  - ✓ **対訳コーパスは主要な言語では大規模に利用できる**
- 機械翻訳におけるドメイン適応 [Chu+ 2018]
- 対話応答生成における転移学習 [Akama+ 2017]
  - → **異なるドメインの大規模パラレルコーパスの存在を仮定**
  - ✓ **本タスクでは状況が違うため生コーパスに基づく手法を提案**
- 教師なしスタイル変換 [Luo+ 2019]
  - → **小規模とは言え、教師あり手法の方が顕著に高い性能を達成**
  - ✓ **大規模な生コーパスと小規模なパラレルコーパスを組み合わせて高品質なスタイル変換を実現**

まとめ：スタイル変換のための折り返し翻訳に基づく事前訓練

- カジュアル → フォーマルの言い換えにおける少資源問題を解消
- 疑似データ（生コーパス）での事前訓練 + 真のデータでの再訓練
- スタイル, ドメイン, モデル構造に依存せず常に大幅に性能改善
- 真のデータが 1k 文対しかない状況でも高品質な言い換えを実現

今後の課題

- データが増えても解けない問題を見つける
- 価値ある疑似データの性質を明らかにする*1
  - 今のところ、自己符号化 < 逆翻訳 < 折り返し翻訳
  - 真のデータに不足している情報を疑似データに入れる*2

*1 Edunov et al. (EMNLP-2018)
Understanding Back-Translation at Scale.
機械翻訳データ拡張の逆翻訳では、ビームサーチよりもサンプリングが良い

*2 Fadaee and Monz (EMNLP-2018)
Back-Translation Sampling by Targeting Difficult Words in NMT.
対訳コーパスの低頻度語を積極的に疑似コーパスに含めると良い