言い換えとリランキングに基づくドメイン不適合の緩和

惟高日向[1]，梶原智之[1]，藤田篤[2]，二宮崇[1]

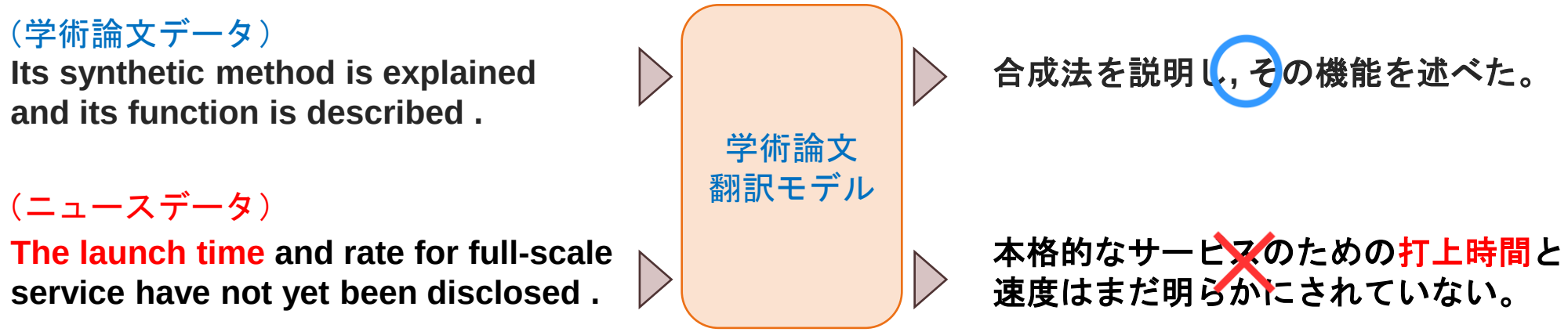
[1]愛媛大学大学院理工学研究科

[2]情報通信研究機構

背景

- 機械翻訳は広く活用されている
  - DeepL
  - Google翻訳 など

- 機械翻訳の品質は訓練データの特徴に依存[1]
  入力文と訓練データの特性が大きく異なる場合には翻訳性能低下の恐れがある[1]



[1] Koehn, P. and Knowles, R. Six Challenges for Neural Machine Translation. In Proc. of NGT 2017.

従来手法とその課題について

- 従来手法: 対象ドメインデータを用いて翻訳モデルを再訓練（fine-tuning）[2]

- 問題点
  1. 対訳コーパスを利用できるドメインが限られている点
  2. 各ドメインのfine-tuningにかかる時間や、構築したモデルの管理コストがかかる点
  3. ブラックボックス化された翻訳器を対象とする場合には適用できない点

[2] Chu, C. and Wang, R. A Survey of Domain Adaptation for Neural Machine Translation. In Proc. of COLING 2018.

関連研究: 機械翻訳における言い換えによる前編集

- 手動前編集:
  言い換えを含む多様なタイプの前編集が翻訳品質を向上させることを示した[3][4]

- 自動前編集
  - 限られた設定（語彙および構文の平易化[5,6,7]）しか検討されていないが品質の向上は確認

➡ 様々な言い換え方法を検討する必要がある

  - 効果的な言い換えの予測ができない[4]

➡ 翻訳品質を事後的に評価する必要がある

[3] Miyata, R. and Fujita, A. Dissecting Human Pre-Editing toward Better Use of Off-the-shelf Machine Translation Systems. In Proc. of EAMT 2017.
[4] Miyata, R. and Fujita, A. Understanding Pre-Editing for Black-Box Neural Machine Translation. In Proc. of EACL 2021.
[5] Stajner, S and Popovic, M. Can Text Simplification Help Machine Translation? In Proc. of EAMT 2016.
[6] Stajner, S and Popovic, M. Improving Machine Translation of English Relative Clauses with Automatic Text Simplification. In Proc. of INLG 2018.
[7] Mehta, S et al. Simplify-Then-Translate: Automatic Preprocessing for Black-Box Translation. In Proc. of AAAI 2020.

提案手法

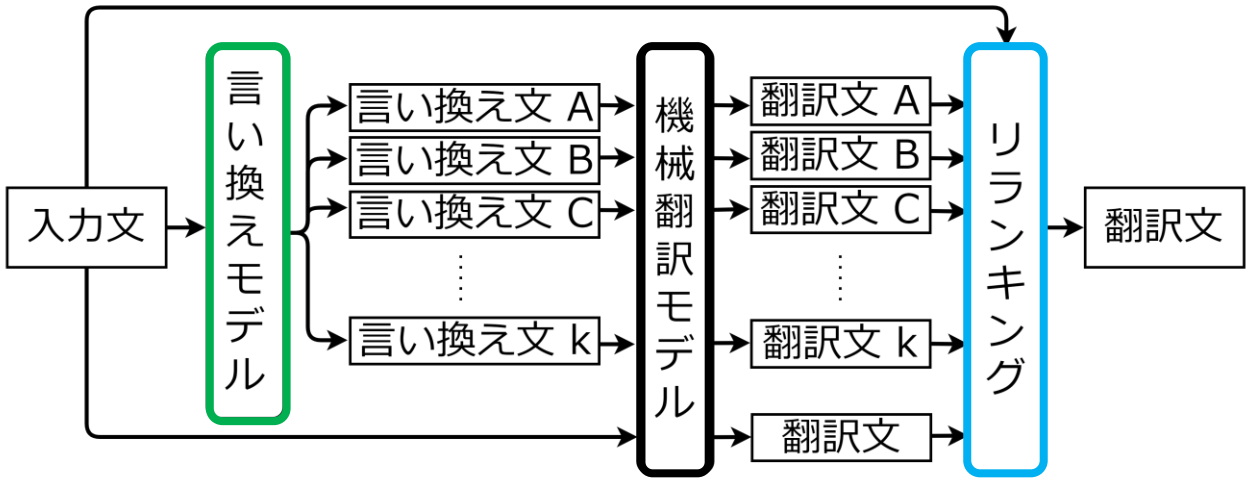
提案手法: 言い換えとリランキングに基づく手法

- 目的: 翻訳対象文とMTの訓練データとの間のドメインギャップを埋める
- アプローチ: **言い換え**(前編集) と**リランキング**(後編集)

1. **言い換え:**
   複数の言い換え文を生成

2. **機械翻訳:**
   入力文と言い換え文を翻訳

3. **リランキング:**
   最良の翻訳文を選択



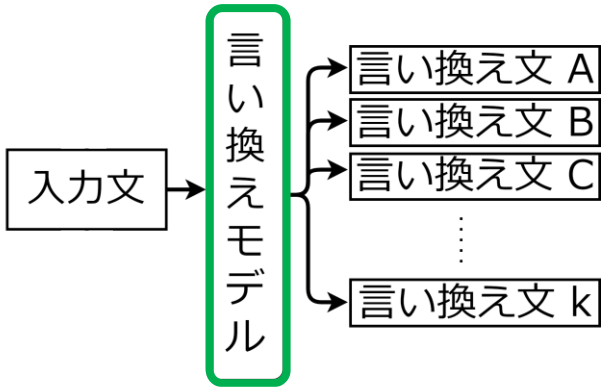
提案手法: 言い換えとリランキングに基づく手法

- 目的: 翻訳対象文とMTの訓練データとの間のドメインギャップを埋める
- アプローチ: **言い換え**(前編集) と**リランキング**(後編集)

1. **言い換え:**
   複数の言い換え文を生成

2. **機械翻訳:**
   入力文と言い換え文を翻訳

3. **リランキング:**
   最良の翻訳文を選択



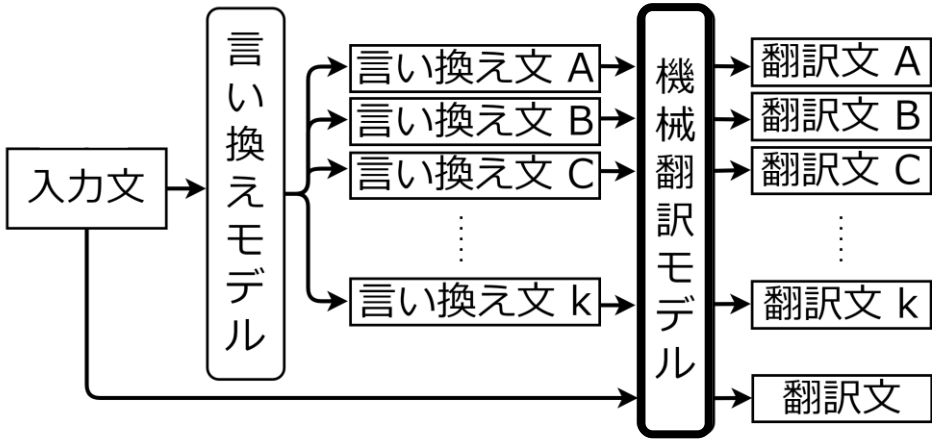
提案手法: 言い換えとリランキングに基づく手法

- 目的: 翻訳対象文とMTの訓練データとの間のドメインギャップを埋める
- アプローチ: **言い換え**(前編集) と**リランキング**(後編集)

1. 言い換え:
   複数の言い換え文を生成

2. **機械翻訳:**
   入力文と言い換え文を翻訳

3. リランキング:
   最良の翻訳文を選択



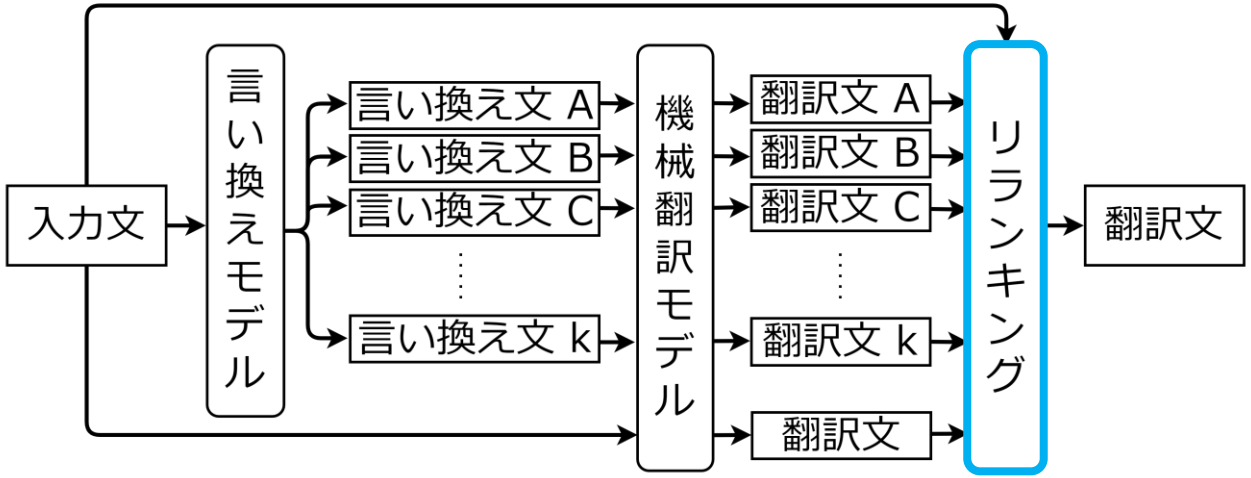
提案手法: 言い換えとリランキングに基づく手法

- 目的: 翻訳対象文とMTの訓練データとの間のドメインギャップを埋める
- アプローチ: **言い換え**(前編集) と**リランキング**(後編集)

1. 言い換え:
   複数の言い換え文を生成

2. 機械翻訳:
   入力文と言い換え文を翻訳

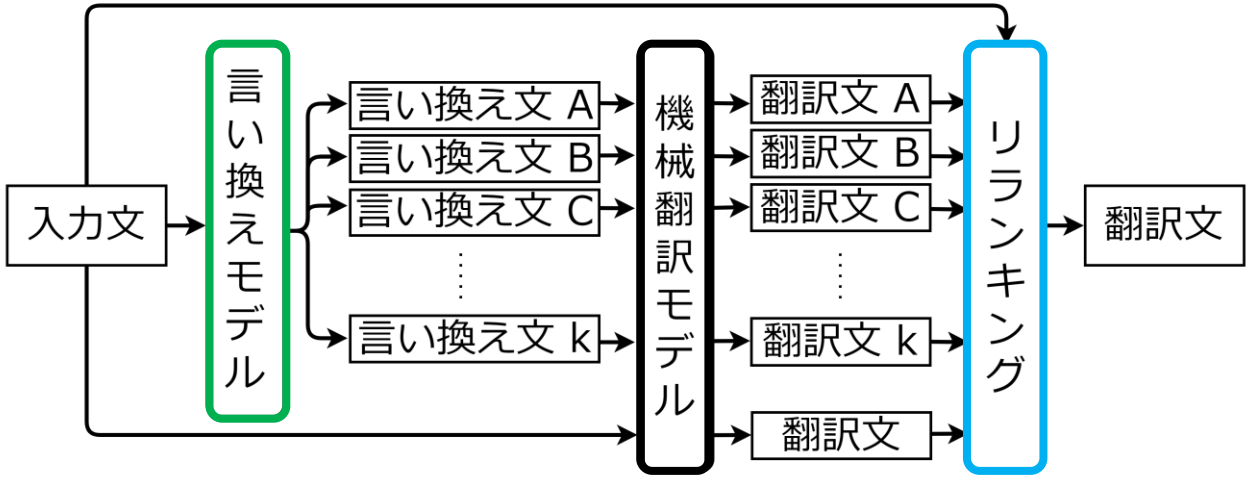
3. **リランキング:**
   最良の翻訳文を選択



言い換え手法: 文単位と単語単位

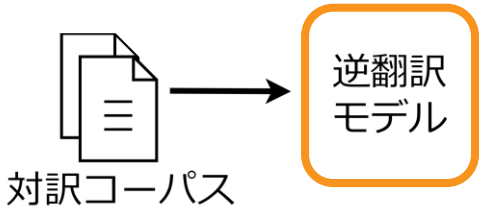
- 文単位の言い換え（既存研究）
  seq2seqモデルを用いて文全体の言い換えを行う
  （逆翻訳に基づくアプローチと単言語翻訳に基づくアプローチ）

- 単語単位の言い換え（提案手法）
  マスク言語モデルと単語埋め込みに基づいて単一単語の言い換えを行う



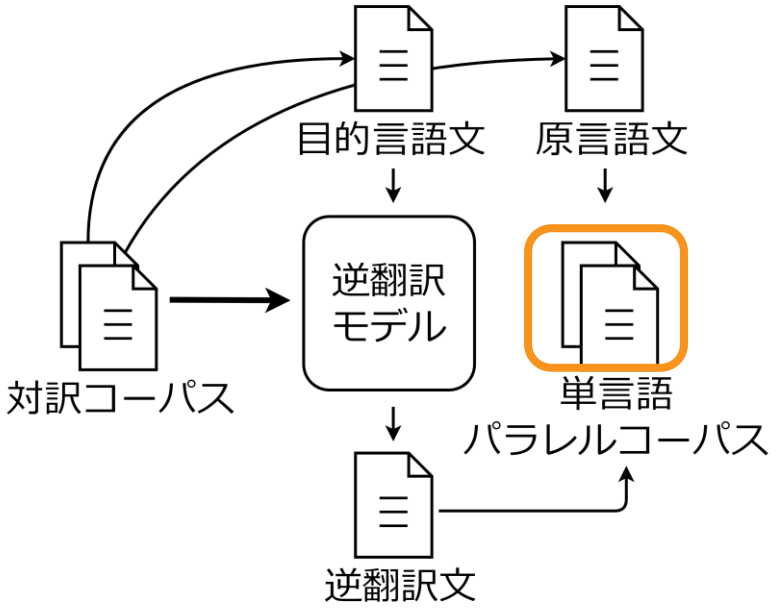
文単位の言い換え: 逆翻訳に基づくアプローチ

1. 逆翻訳モデルの訓練
2. 対訳コーパスの目的言語側を原言語側に翻訳し、原言語側と結合
3. 作成した単言語コーパスを使って言い換えモデルを訓練



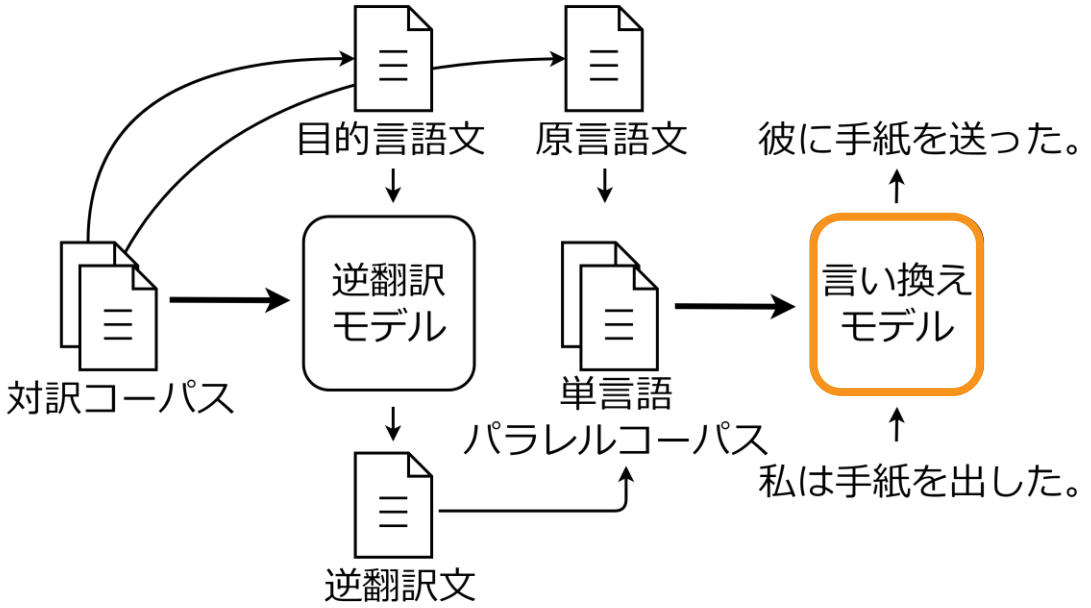
文単位の言い換え: 逆翻訳に基づくアプローチ

1. 逆翻訳モデルの訓練
2. 対訳コーパスの目的言語側を原言語側に翻訳し、原言語側と結合
3. 作成した単言語コーパスを使って言い換えモデルを訓練



文単位の言い換え: 逆翻訳に基づくアプローチ

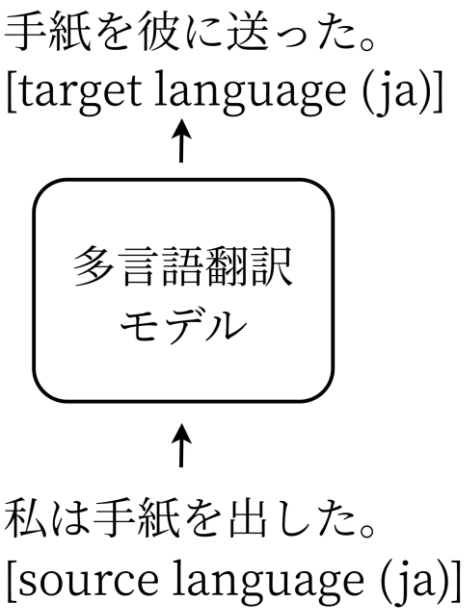
1. 逆翻訳モデルの訓練
2. 対訳コーパスの目的言語側を原言語側に翻訳し、原言語側と結合
3. 作成した単言語コーパスを使って言い換えモデルを訓練



文単位の言い換え: 単言語翻訳に基づくアプローチ

事前訓練された多言語翻訳モデルを使用して、文単位の言い換えを生成

- mBARTやm2mのような原言語をカバーできるモデル
- 原言語と目的言語に同じ言語を指定

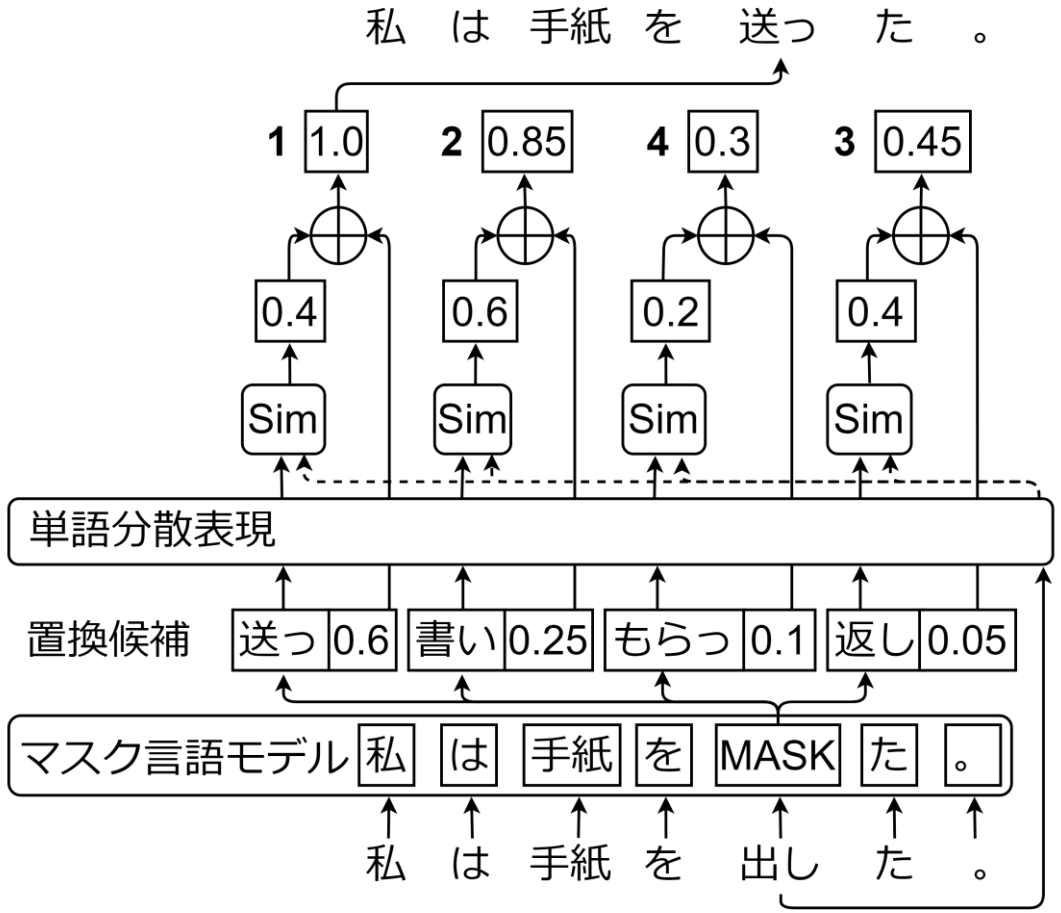


単語単位の言い換え

1. 置換候補の生成 (|X|×n)
   - 各単語x_iがマスクされた文をMLMに入力
   - 各単語n個の置換候補を生成

2. トップk個の置換候補の選択

   以下の数値の和で候補をリランキング
   - MLMの生成確率
   - 元単語x_iと言い換え候補の単語埋め込みの余弦類似度



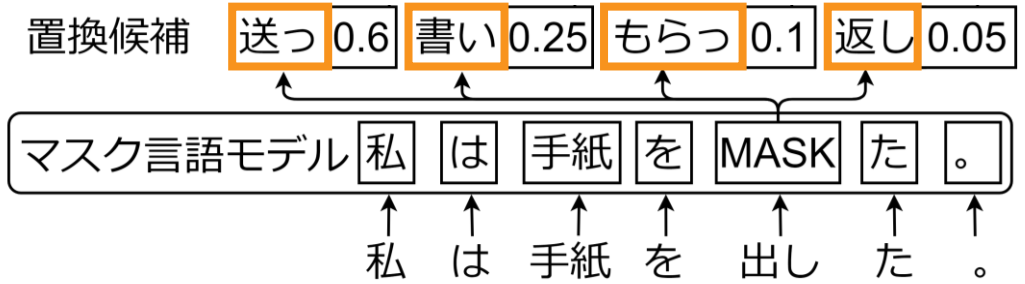
|X|個の単語から成る入力文: $X = x_1, \ldots, x_X$

単語単位の言い換え: 置換候補の生成

1. 置換候補の生成 ($|X|$×n)
   - 各単語x_iがマスクされた文をMLMに入力
   - 各単語n個の置換候補を生成

2. トップk個の置換候補の選択

   以下の数値の和で候補をリランキング
   - MLMの生成確率
   - 元単語x_iと言い換え候補の単語埋め込みの余弦類似度



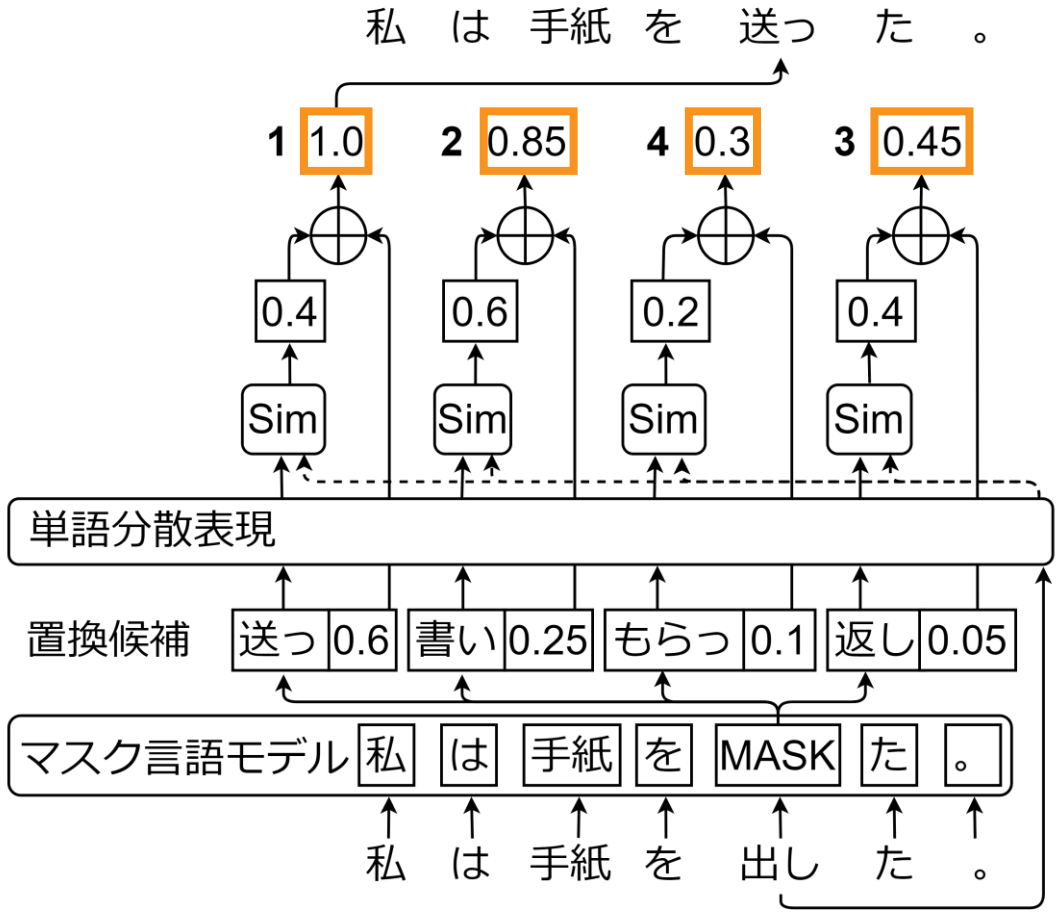
$|X|$個の単語から成る入力文: $X = x_1, \ldots, x_X$

単語単位の言い換え: 置換候補の選択

1. 置換候補の生成 (|X|×n)
   - 各単語x_iがマスクされた文をMLMに入力
   - 各単語n個の置換候補を生成

2. トップk個の置換候補の選択

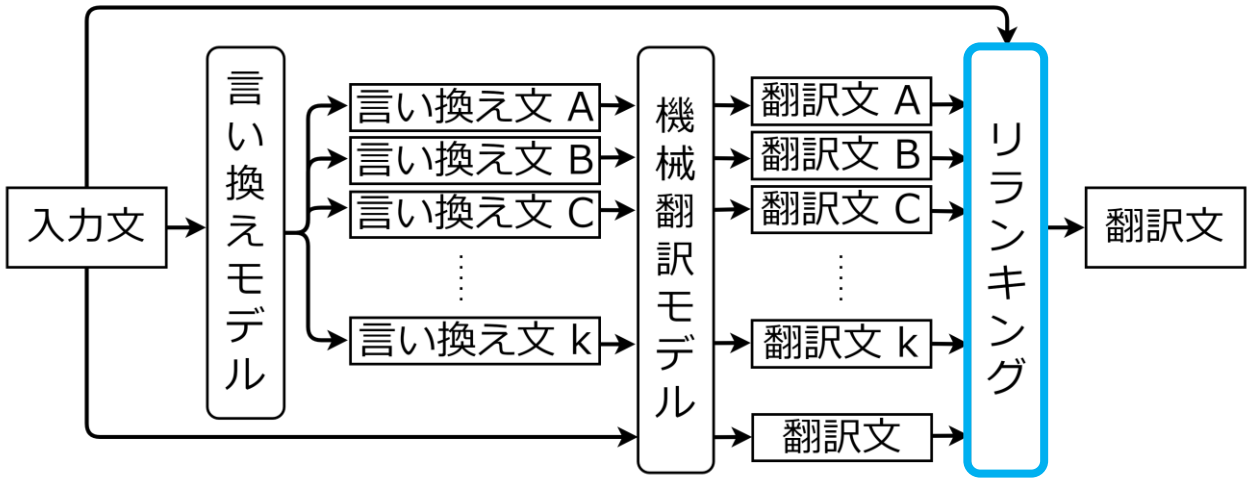
   以下の数値の和で候補をリランキング
   - MLMの生成確率
   - 元単語x_iと言い換え候補の単語埋め込みの余弦類似度



|X|個の単語から成る入力文: $X = x_1, \ldots, x_X$

リランキング手法

1. 入力文と各翻訳候補の双方向でforced decoding[8,9]を計算
2. 双方向のforced decodingスコアの平均によって翻訳候補をリランキング



- Black-box設定: 原言語・目的言語をカバーする多言語翻訳モデルを使用
- Glass-box設定: 対象翻訳モデルとその逆方向の翻訳モデルを使用

[8] Marie, B. and Fujita, A. A Smorgasbord of Features to Combine Phrase-Based and Neural Machine Translation. In Proc. of AMTA 2018.
[9] Kiyono, S et al. Tohoku-AIP-NTT at WMT 2020 news translation task. In Proc. of WMT 2020.

日英機械翻訳の実験

実験設定: データセット

	Corpus	Domain	#lines
訓練 (翻訳モデル)	JParaCrawl[10]	General	10,000,000
検証 (翻訳モデル)	JParaCrawl	General	2,000
訓練 (言い換えモデル)	JParaCrawl	General	10,000,000
検証 (言い換えモデル)	JParaCrawl	General	2,000
評価	ASPEC[11]	科学技術論文	1,812
評価	WMT20[12]	ニュース	993
評価	MTNT2019[13]	Reddit	1,033
評価	JParaCrawl	General	2,000

[10] Morishita, M et al. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In Proc. of LREC 2022.
[11] Nakazawa, T et al. ASPEC: Asian Scientific Paper Excerpt Corpus. In Proc. of LREC 2016.
[12] Barrault, L et al. Findings of the 2020 Conference on Machine Translation. In Proc. of WMT 2020.
[13] Li, X et al. Findings of the first shared task on machine translation robustness. In Proc. of WMT 2019.

実験設定: モデル

- 機械翻訳: Transformer
- 文単位の言い換え
  - 逆翻訳に基づくアプローチ: Transformer
  - 単言語翻訳に基づくアプローチ: mBART or M2M-100
- 単語単位の言い換え
  - マスク言語モデル: BERT
  - 単語埋め込み: fasttext
- Black-boxリランキング: mBART

[14] Vaswani, A et al. Attention is All you Need. In Proc. of NIPS 2017.
[15] Tang, Y et al. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. CoRR 2020.
[16] Fan, A et al. Beyond English-Centric Multilingual Machine Translation. Journal of Machine Learning Research 2021.
[17] Devlin, J et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proc. of NAACL 2019.
[18] Bojanowski, P et al. Enriching Word Vectors with Subword Information. TACL 2017.

実験設定:

- 言い換えを使わない比較手法（k: 言い換えの数）
  - ➢beam探索: beam幅($k+1$)のビーム探索を用いて1bestを出力
  - ➢beam探索 + リランキング: beam幅($k+1$)のbeam探索の後に、リランキングを行う

- デコード手法: beam探索

- 評価指標: BLEU

実験結果: 言い換え手法に着目

- ASPECとWMT20: 単語単位の言い換え手法が文単位よりも常に良い性能
- MTNT2019とJParaCrawl: 単語単位と文単位の言い換え手法に優劣な差はない

	言い換え		リランキング	ASPEC			WMT20			MTNT2019			JParaCrawl		
ID	単位	モデル	モデル	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
(1)	-	-	-	20.5	20.7	20.6	20.6	20.9	20.8	15.5	15.6	15.9	34.6	34.6	34.5
(2)	-	-	mBART (black-box)	20.6	20.8	21.1*	20.8	21.2*	21.4*	**15.8**	**15.9**	**15.9**	**34.1***	**33.9***	**33.4***
(3)	単語	JaBERT		20.9*	21.0	20.8	**21.5***	**21.8***	**21.9***	15.6	15.6	15.5	33.3*	32.9*	32.6*
(4)	単語	mBERT		**21.1***	**21.2***	**21.2***	21.2*	21.4*	21.8*	15.4	15.3	15.4	33.3*	33.1*	32.7*
(5)	文	mBART		20.3	20.1*	20.1*	20.3	20.4*	20.4*	14.8*	14.6*	14.7*	33.1*	32.9*	32.2*
(6)	文	M2M-100		19.9*	19.8*	19.7*	20.8	20.6	20.8	15.3	15.2	14.5*	32.9*	32.5*	32.0*
(7)	文	Denoiser		20.4	20.3*	20.4	20.7	20.9	20.9	15.5	15.3	15.4	33.8*	33.1*	32.9*
(8)	-	-	MT (glass-box)	20.7	21.1*	21.2*	20.8	20.9	21.2*	15.8*	**15.9**	**15.9**	**34.7**	**34.7**	34.3
(9)	単語	JaBERT		**21.2***	**21.4***	21.2*	**21.6***	**22.0***	21.9*	**15.9***	15.7	15.5	34.5	34.5	34.3
(10)	単語	mBERT		**21.2***	21.3*	**21.6***	**21.6***	21.8*	**22.1***	15.7	15.4	15.5	34.5	34.4	34.0*
(11)	文	mBART		20.8*	20.7	21.0*	21.0*	20.9	21.1	15.4	15.5	15.4	**34.7**	34.6	34.6
(12)	文	M2M-100		20.6	20.6	20.6	21.2*	21.4*	21.5*	15.8	**15.9**	15.7	34.6	**34.7**	**34.7**
(13)	文	Denoiser		20.6	20.7	20.8	20.9*	20.9	21.2*	15.8	15.7	15.8	**34.7**	**34.7**	**34.7**

BLEU スコア(k: 言い換えの数, 太字: 各リランキング結果における列ごとの最高BLEUスコア,
*: 一行目のベースライン手法と比較して統計的に有意($p < 0.05$))

実験結果: 言い換えの数に着目

- ASPECとWMT20: 単語単位ではkが大きくなるにつれて翻訳品質が向上
- MTNT2019とJParaCrawl: kが増加しても翻訳品質は低下もしくは変化なし

	言い換え		リランキング	ASPEC			WMT20			MTNT2019			JParaCrawl		
ID	単位	モデル	モデル	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$	$k=5$	$k=10$	$k=20$
(1)	-	-	-	20.5	20.7	20.6	20.6	20.9	20.8	15.5	15.6	15.9	34.6	34.6	34.5
(2)	-	-	mBART (black-box)	20.6	20.8	21.1*	20.8	21.2*	21.4*	**15.8**	**15.9**	**15.9**	**34.1***	**33.9***	**33.4***
(3)	単語	JaBERT		20.9*	21.0	20.8	**21.5***	**21.8***	**21.9***	15.6	15.6	15.5	33.3*	32.9*	32.6*
(4)	単語	mBERT		**21.1***	**21.2***	**21.2***	21.2*	21.4*	21.8*	15.4	15.3	15.4	33.3*	33.1*	32.7*
(5)	文	mBART		20.3	20.1*	20.1*	20.3	20.4*	20.4*	14.8*	14.6*	14.7*	33.1*	32.9*	32.2*
(6)	文	M2M-100		19.9*	19.8*	19.7*	20.8	20.6	20.8	15.3	15.2	14.5*	32.9*	32.5*	32.0*
(7)	文	Denoiser		20.4	20.3*	20.4	20.7	20.9	20.9	15.5	15.3	15.4	33.8*	33.1*	32.9*
(8)	-	-	MT (glass-box)	20.7	21.1*	21.2*	20.8	20.9	21.2*	15.8*	**15.9**	**15.9**	**34.7**	**34.7**	34.3
(9)	単語	JaBERT		**21.2***	**21.4***	21.2*	**21.6***	**22.0***	21.9*	**15.9***	15.7	15.5	34.5	34.5	34.3
(10)	単語	mBERT		**21.2***	21.3*	**21.6***	**21.6***	21.8*	**22.1***	15.7	15.4	15.5	34.5	34.4	34.0*
(11)	文	mBART		20.8*	20.7	21.0*	21.0*	20.9	21.1	15.4	15.5	15.4	**34.7**	34.6	34.6
(12)	文	M2M-100		20.6	20.6	20.6	21.2*	21.4*	21.5*	15.8	**15.9**	15.7	34.6	**34.7**	**34.7**
(13)	文	Denoiser		20.6	20.7	20.8	20.9*	20.9	21.2*	15.8	15.7	15.8	**34.7**	**34.7**	**34.7**

BLEU スコア(k: 言い換えの数, 太字: 各リランキング結果における列ごとの最高BLEUスコア,
*: 一行目のベースライン手法と比較して統計的に有意($p < 0.05$))

実験結果: 言い換え手法の組み合わせ

単語単位の手法と比較:
　ほとんどの組み合わせでBLEUスコアは低下・もしくは変化なし

ID	言い換え 単位	言い換え モデル	リランキング モデル	ASPEC $k=5$	ASPEC $k=10$	ASPEC $k=20$	WMT20 $k=5$	WMT20 $k=10$	WMT20 $k=20$
(1)	-	-	-	20.7	20.6	20.8	20.9	20.8	20.7
(2)	-	-	mBART (black-box)	**20.8**	**21.1***	**21.2***	**21.2***	21.4*	21.2*
(3)	単語 + 文	JaBERT + mBART		20.5	20.5	20.7	21.1	21.5*	21.4*
(4)	単語 + 文	JaBERT + M2M-100		20.2*	20.4	20.2*	**21.2**	21.4*	21.5*
(5)	単語 + 文	JaBERT + Denoiser		20.7	20.8	20.8	**21.2**	**21.7***	**21.7***
(6)	単語 + 文	mBERT + mBART		20.6	20.5	20.9	21.0	21.3*	21.5*
(7)	単語 + 文	mBERT + M2M-100		20.5	20.6	20.7	21.1	21.2	21.5*
(8)	単語 + 文	mBERT + Denoiser		**20.8**	20.8	20.8	21.0	21.2	21.6*
(9)	-	-	MT (glass-box)	21.1*	21.2*	21.3*	20.9	21.2*	21.0
(10)	単語 + 文	JaBERT + mBART		**21.2***	21.2*	21.3*	21.7*	21.9*	21.7*
(11)	単語 + 文	JaBERT + M2M-100		**21.2***	**21.3***	21.1	**21.8***	**22.2***	21.9*
(12)	単語 + 文	JaBERT + Denoiser		21.0	**21.3***	21.2*	21.5*	22.0*	21.8*
(13)	単語 + 文	mBERT + mBART		**21.2***	21.2*	**21.6***	21.6*	21.8*	21.9*
(14)	単語 + 文	mBERT + M2M-100		**21.2***	21.2*	21.4*	**21.8***	21.9*	**22.1***
(15)	単語 + 文	mBERT + Denoiser		21.1*	**21.3***	21.4*	21.5*	21.7*	**22.1***

BLEU スコア(k: 単語単位と文単位のそれぞれの言い換えの数, 太字: 各リランキング結果における列ごとの最高BLEUスコア, *: 一行目のベースライン手法と比較して統計的に有意($p < 0.05$)
下線: 単語単位の言い換えと比較して改善したBLEUスコア

分析

分析: Oracleに着目

単語単位の言い換えの値が最も大きい

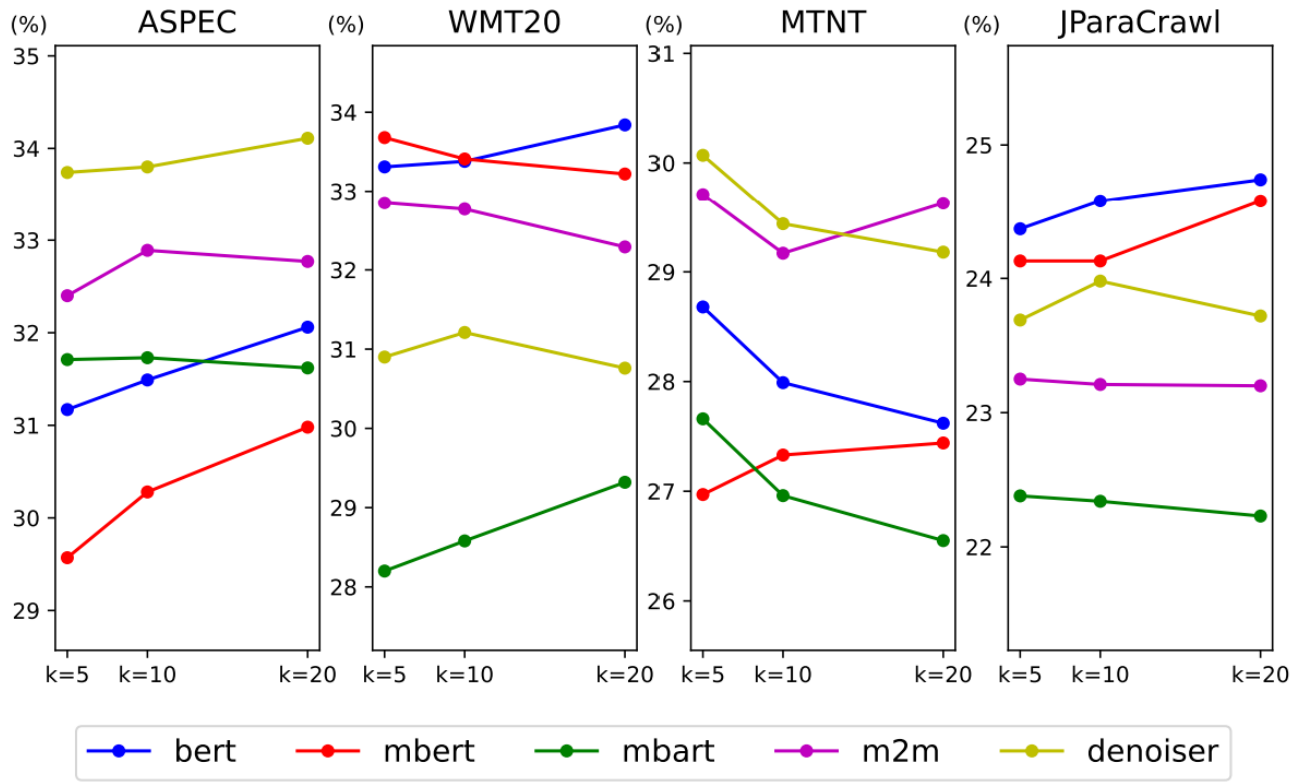
文単位の言い換えや言い換えなしのビーム探索と比較して、翻訳性能改善のポテンシャルが高い

ID	言い換え 単位	言い換え モデル	リランキング モデル	ASPEC $k=5$	ASPEC $k=10$	ASPEC $k=20$	WMT20 $k=5$	WMT20 $k=10$	WMT20 $k=20$	MTNT2019 $k=5$	MTNT2019 $k=10$	MTNT2019 $k=20$	JParaCrawl $k=5$	JParaCrawl $k=10$	JParaCrawl $k=20$
(14)	-	-	Oracle	**24.3**	26.0	**27.9**	23.6	25.3	26.8	19.0	20.2	**22.0**	**38.6**	**40.1**	**41.5**
(15)	単語	JaBERT		**24.3**	25.9	27.4	24.9	**26.3**	27.6	**19.3**	**20.4**	21.8	37.8	38.7	39.8
(16)	単語	mBERT		**24.3**	**26.1**	**27.9**	**25.0**	26.2	**27.7**	19.0	**20.4**	21.7	37.6	38.6	39.7
(17)	文	mBART		23.4	24.4	25.4	23.1	23.8	24.8	17.8	18.3	19.2	36.4	37.1	37.6
(18)	文	M2M-100		23.6	24.5	25.5	23.8	24.5	25.5	18.2	19.0	19.8	36.6	37.1	37.7
(19)	文	Denoiser		23.5	24.6	25.8	23.7	24.7	25.4	18.4	19.3	20.3	36.7	37.4	38.1

BLEU スコア(k: 言い換えの数, 太字: 各リランキング結果における列ごとの最高BLEUスコア,
*: 一行目のベースライン手法と比較して統計的に有意($p < 0.05$))

分析: 言い換え文の翻訳文に着目

- より良い言い換えが翻訳改善を必ずしも保証はしない
- より良い翻訳候補を得るためにkを増やすことが有効な場合もある
- リランキング手法のさらなる改善が必要



まとめ

- 背景:
  既存のアプローチでは、多くのドメインをカバーするには限界があった
- 目的
  入力文と翻訳器の訓練データのドメインのギャップを埋める
- 手法
  言い換え・機械翻訳・リランキングから成るフレームワークを提案
- 結果
  特に、単語単位とグラスボックス・リランキングを組み合わせることで、
  2つのドメインにおいて一貫して翻訳品質が向上