

事前訓練済みモデル・LLMを用いた 特許翻訳の二段階自動後編集

○武馬 光星¹ 西村 柁人² 宇津呂 武仁² 永田昌明³

1. 筑波大学 理工学群 工学システム学類 自然言語処理研究室
2. 筑波大学 システム情報工学研究群 自然言語処理研究室
3. NTTコミュニケーション科学基礎研究所

- 背景
- 関連研究
- 提案法
- 実験と結果
- まとめ

- 背景
- 関連研究
- 提案法
- 実験と結果
- まとめ

- **翻訳文の後編集とは**

- 機械翻訳で生成された訳文を人間が修正して、より自然で正確な訳に仕上げるプロセス

- **本研究では**

- 翻訳誤り検出に事前訓練済みモデル(mBERT)を、翻訳誤り訂正に大規模言語モデル(GPT-4o)を用いることでこのプロセスを自動化
- 特許文の翻訳に対して適用

1. **誤りの検出**

日本語原文：私は白い猫と犬が好きです。
英訳：I like **black** cats and **birds**.



2. **誤りの訂正**

修正後の英訳：I like **white** cats and **dogs**.

● 目的

- 書き手の翻訳作業の負担軽減
- 特に、日本語の特許を海外で出願するためなど、誤りが許されない翻訳を行う必要がある場合
- 誤り箇所を挙げてその部分を訂正することで、書き手の確認する文量（作業量）を大きく低減する
- 比較
 - ・ 一般的な機械翻訳 … 翻訳後の文の翻訳精度が高いことが重要
 - ・ 本研究 … 誤り検出の精度(特にrecall)が高いかつ、その箇所を訂正することが重要

特許文は、従来の翻訳モデルでの翻訳が難しい

- 専門性が高く、厳密な表現や文構造を必要とする
- 特に、特許請求項では 1 文が極めて長い

・ 特許請求項の 1 文の例

回転機構を備える生産設備により生じる振動を検出するセンサが出力する振動データを取得する、取得部と、前記生産設備の設置時、且つ、負荷がない状態で前記生産設備が稼働された状態で取得された前記振動データから計算した周波数スペクトルを初期波形として記憶し、負荷がない状態で前記生産設備が稼働された状態で取得された前記振動データから計算した周波数スペクトルを始動波形として記憶し、負荷がある状態で前記生産設備が稼働された状態で取得された前記振動データから計算した周波数スペクトルを動作波形として記憶する、FFT(Fast Fourier Transform)部と、前記始動波形及び前記動作波形に基づき、前記生産設備による生産材の加工が正常か否かを判定する、判定部と、を備え、前記判定部は、前記生産材の加工が正常か否かを判定する際に用いる閾値を、前記初期波形及び前記始動波形に基づき補正する、診断装置。

・ 原文の一部が翻訳されない翻訳ミス

入力文

食品総量100質量部に対して、上記の組成物は、0.01～15質量部であり、好ましくは食品総量100質量部に対して、上記の組成物は、0.05～10質量部であり、好ましくは食品総量100質量部に対して、上記の組成物は、0.1～5質量部であり、および/または、上記の食品は、ココアバターおよび/またはココアバターのようなものおよび/またはココアバター代用脂を含有する食品であり、および/または、食品総量100質量部に対して、さらにココアバターおよび/またはココアバターのようなものおよび/またはココアバター代用脂30～50質量部を含有し、および/または、食品総量100質量部に対して、さらにココアパウダー10～20質量部、糖30～50質量部、リン脂質0.1～5質量部を含有し、および/または、上記の食品はチョコレートまたはチョコレートを含有する食品である、請求項6に記載された食品。

参照訳文

The food of claim 6, wherein with respect to 100 parts by weight of the total weight of the food, the composition is 0.01-15 parts by weight; preferably with respect to 100 parts by weight of the total weight of the food, the composition is 0.05-10 parts by weight; more preferably with respect to 100 parts by weight of the total weight of the food, the composition is 0.1-5 parts by weight; and/or, the food is a food comprising cocoa butter and/or a cocoa butter equivalent and/or a cocoa butter replacer; and/or, the food also comprises 30-50 parts by weight of the cocoa butter and/or cocoa butter equivalent and/or cocoa butter replacer, with respect to 100 parts by weight of the total weight of the food; and/or, the food also comprises 10-20 parts by weight of cocoa powder, 30-50 parts by weight of sugar and 0.1-5 parts by weight of phospholipid, with respect to 100 parts by weight of the total weight of the food; and/or, the food is chocolate or a food comprising chocolate.

NMTモデルの出力文

The food product according to claim 6, wherein the amount of the composition is 0.01 to 15 parts by mass based on 100 parts by mass of the total amount of the food product, and/or the amount of the composition is 0.05 to 10 parts by mass based on 100 parts by mass of the total amount of the food product, and/or the amount of the composition is 0.1 to 5 parts by mass based on 100 parts by mass of the total amount of the food product.

・余分な文が繰り返されてしまう翻訳ミス

入力文

a.配電ハードウェアの構成部品として、少なくとも1つの受動電磁センサをインストールするステップと、 b.

参照訳文

a.Installing at least one Passive Electromagnetic Sensor as a component of distribution hardware;b.

NMTモデルの出力文

a. installing at least one passive electromagnetic sensor as a component part of the electrical distribution hardware;b. controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;c. controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;d. controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;e. controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;f. controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;g. controlling the at least one passive electromagnetic sensor to emit electromagnetic radiation;g.

- 背景
- **関連研究**
- 提案法
- 実験と結果
- まとめ

単語にOK/BADタグを付与することで、翻訳品質評価を行う

- 原言語側の挿入、置換が必要なトークンにBADタグ
- 目的言語側の削除、置換が必要なトークンにBADタグ
- 正しい翻訳のトークンにOKタグ

原言語文：なお、変化率 K_i は定数である。

目的言語文：The temperature change rate K_i of a sample .



翻訳後に翻訳文の訂正を行うことで、品質を向上させる

- 機械翻訳の品質向上に向けて、後編集において LLM を活用する
- 原文、翻訳文と共に様々な粒度の情報をLLMに入力し、LLMで翻訳訂正

原文 : Naturally **tanned** ... dye or chemicals.

翻訳 : Natürlich ... verdünnt ... oder Chemikalien.



情報 1 : The translation is 95 out of 100.

翻訳の点数

情報 2 : The translation contains a major mistranslation error at “verdünnt”.

翻訳エラー



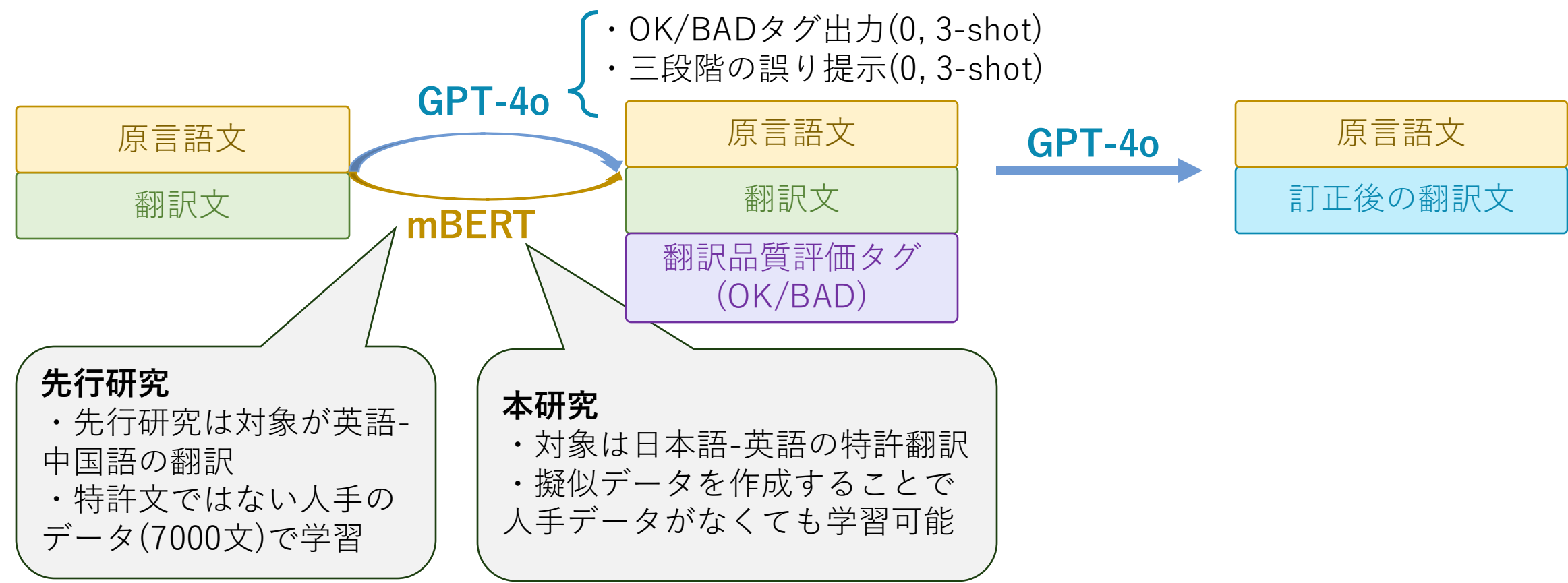
修正後の訳 : Natürlich **gegerbt** nur ... oder Chemikalien.

- 背景
- 関連研究
- **提案法**
- 実験と結果
- まとめ

● 翻訳誤り検出 + 翻訳誤り訂正の2段階の手法

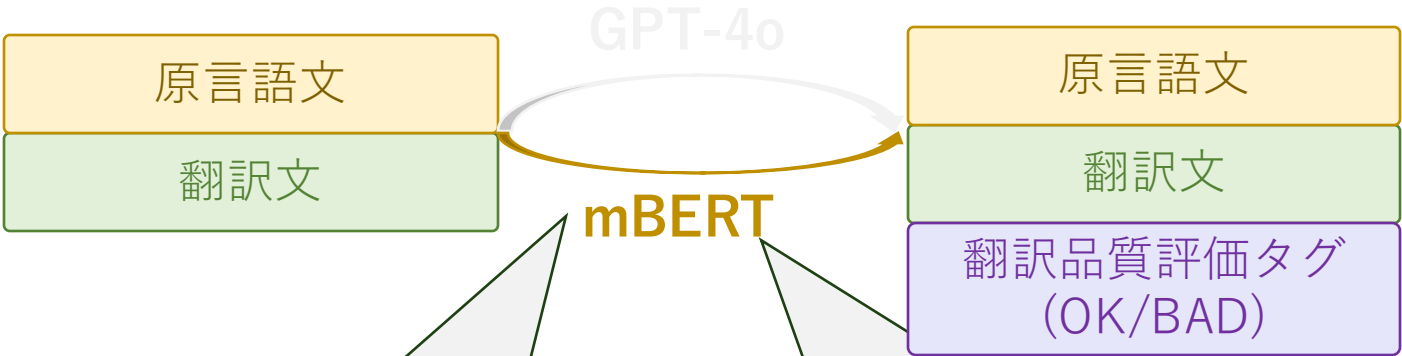
1. 翻訳誤り検出

2. 翻訳誤り訂正



- 翻訳誤り検出 + 翻訳誤り訂正の 2 段階の手法

1. 翻訳誤り検出



先行研究

- ・ 先行研究は対象が英語-中国語の翻訳
- ・ 特許文ではない人手のデータ(7000文)で学習

本研究

- ・ 対象は日本語-英語の特許翻訳
- ・ 擬似データを作成することで人手データがなくても学習可能

・ 誤り検出モデル(mBERT)の学習を行うために、擬似データを作成

・ 先行研究を参考に対訳特許文に以下の手法で人工的な誤りを発生させる

1. トークンを5%の確率で**削除**
2. トークンを10%の確率で**挿入**
3. トークンを30%の確率で**置換**

例 (置換した場合)

The **change** rate K_i is a constant .

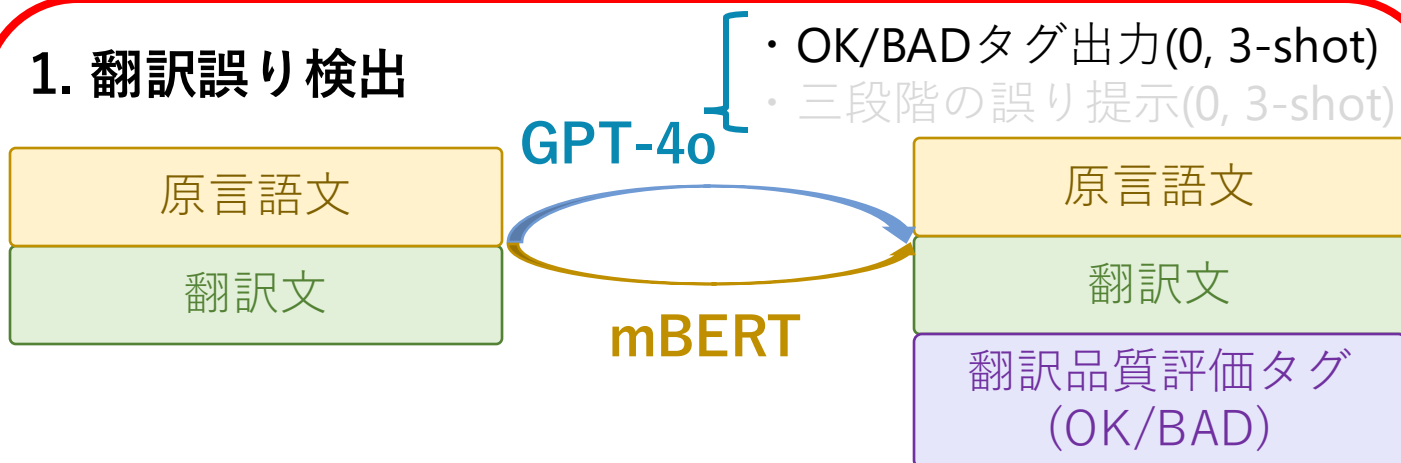


The **challenge** rate K_i is a constant .

OK BAD OK OK OK OK OK OK

- 背景
- 関連研究
- 提案法
- **実験と結果**
- まとめ

1. 翻訳誤り検出



・ 正しく誤りを検出できるか評価するためにOK/BADタグを出力させる

● 学習データ

・ NTCIR 7, NTCIR 8 の特許文対訳データに誤りを作成したデータ：8000文

● テストデータ

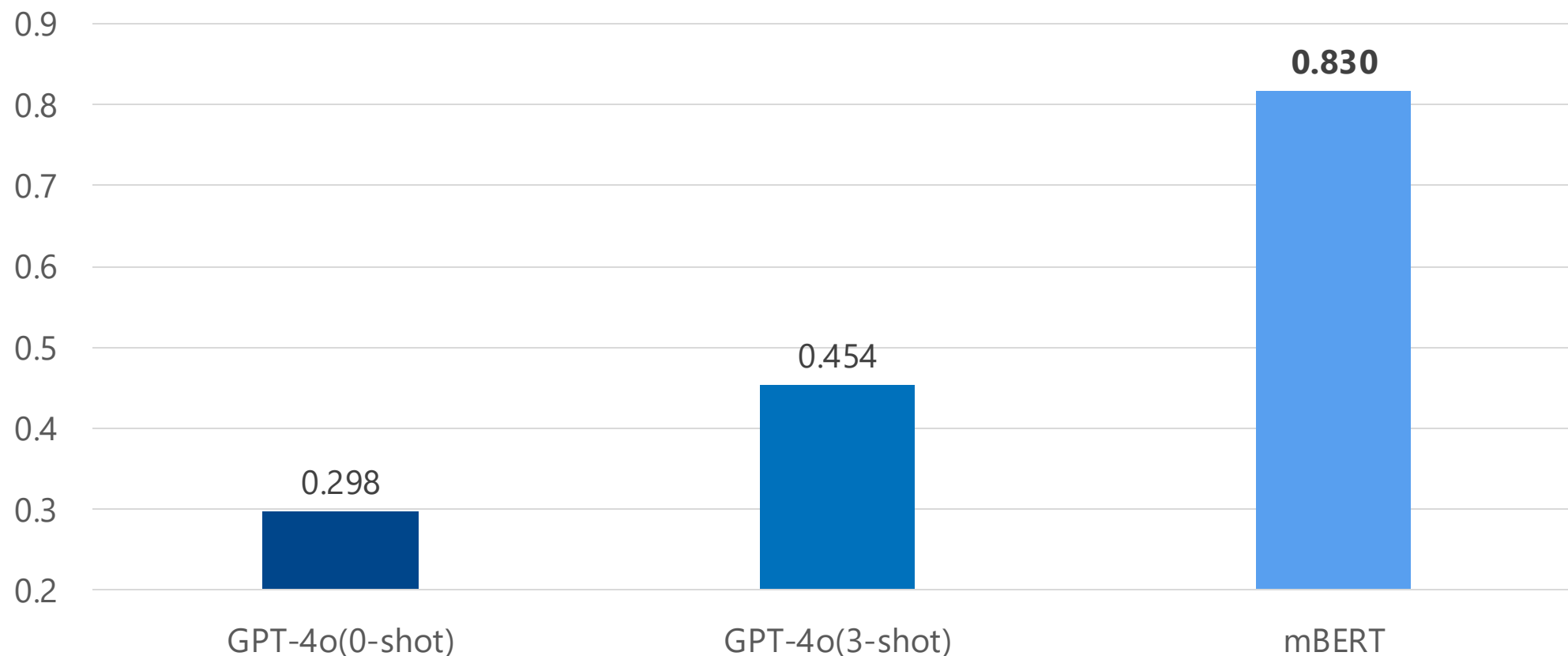
・ NTCIR 7, NTCIR 8 の特許文対訳データに誤りを作成したデータ：1000文

● 評価指標

・ F1値
・ PrecisionとRecallの調和平均で、分類モデルの全体的なバランスを測る指標

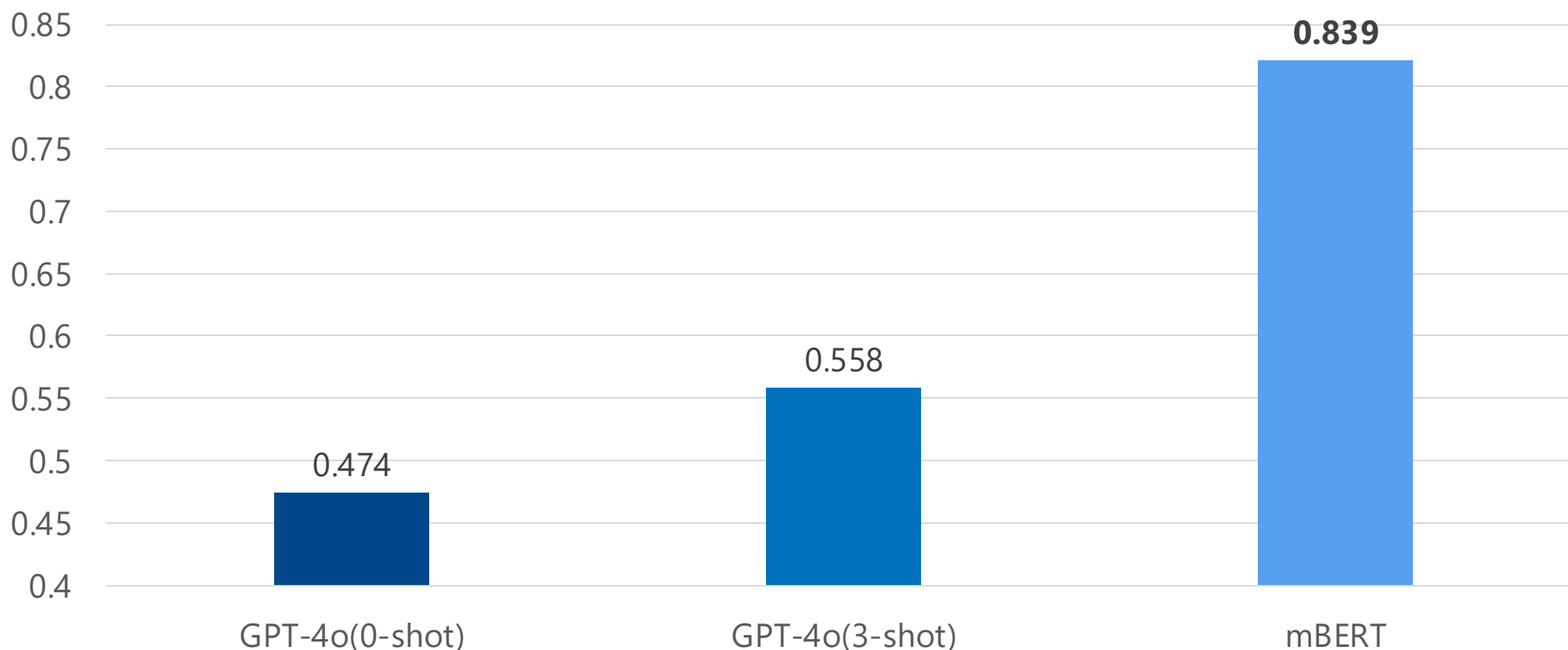
- 特許文で学習を行なったmBERT(提案手法)とGPT-4oの検出精度比較

- ・ 日英翻訳 (F1値)



- 特許文で学習を行なったmBERT(提案手法)とGPT-4oの検出精度比較

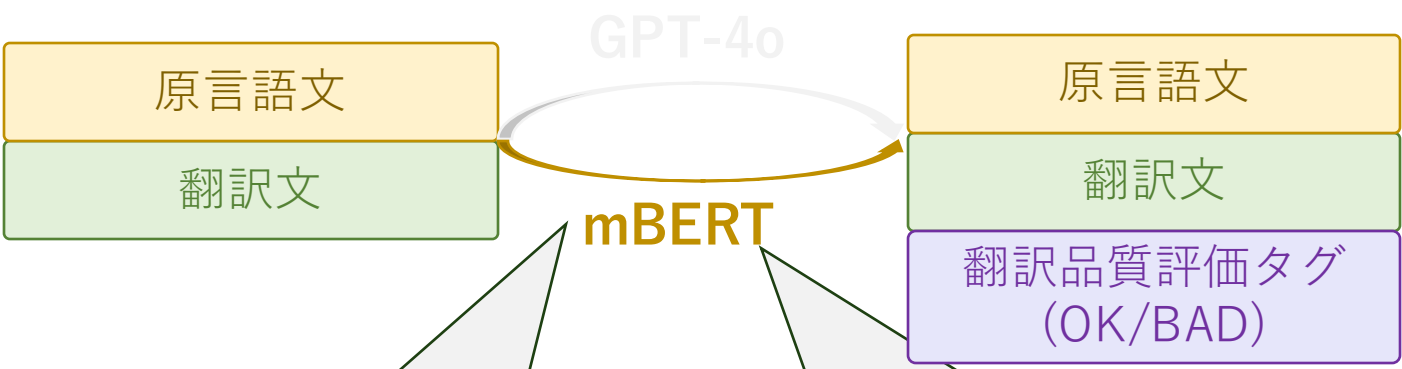
- ・ 英日翻訳 (F1値)



- ・ 様々なタスクで高い精度を誇るLLMを大きく上回る精度を達成

● 翻訳誤り検出 + 翻訳誤り訂正の 2 段階の手法

1. 翻訳誤り検出



先行研究

- ・ 先行研究は対象が英語-中国語の翻訳
- ・ 特許文ではない人手のデータ(7000文)で学習

本研究

- ・ 対象は日本語-英語の特許翻訳
- ・ 人工データを作成することで人手データがなくても学習可能

・ 誤り検出モデル(mBERT)の学習を行うために、擬似データを作成

・ 先行研究を参考に対訳特許文に以下の手法で人工的な誤りを発生させる

1. トークンを5%の確率で**削除**
2. トークンを10%の確率で**挿入**
3. トークンを30%の確率で**置換**

例 (置換した場合)

The **change** rate Ki is a constant .

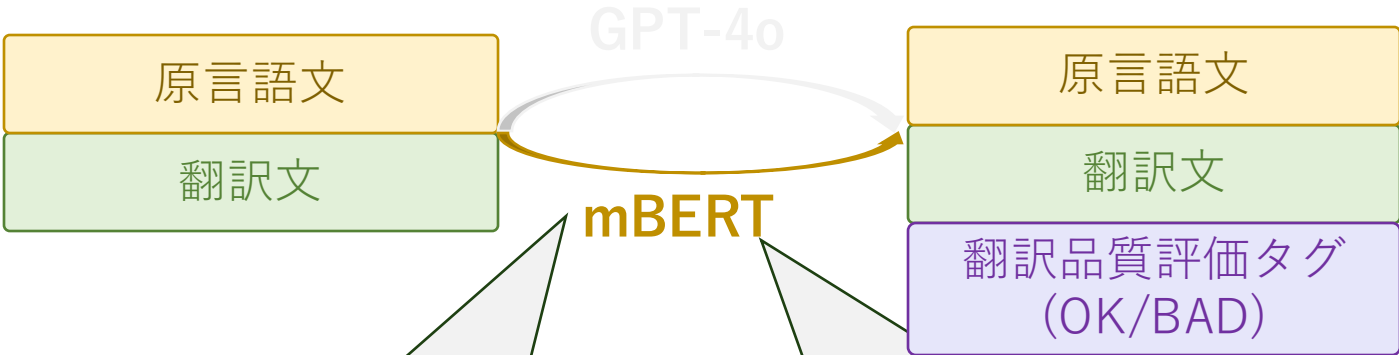
↓ **最も適切でないトークン**

The **challenge** rate Ki is a constant .

OK BAD OK OK OK OK OK OK

- 翻訳誤り検出 + 翻訳誤り訂正の 2 段階の手法

1. 翻訳誤り検出



先行研究

- ・ 先行研究は対象が英語-中国語の翻訳
- ・ 特許文ではない人手のデータ(7000文)で学習

本研究

- ・ 対象は日本語-英語の特許翻訳
- ・ 人工データを作成することで人手データがなくても学習可能

・ 誤り検出モデル(mBERT)の学習を行うために、擬似データを作成

・ 先行研究を参考に対訳特許文に以下の手法で人工的な誤りを発生させる

1. トークンを5%の確率で**削除**
2. トークンを10%の確率で**挿入**
3. トークンを30%の確率で**置換**

例 (置換した場合)

The **change** rate Ki is a constant .



最も確からしいトークン

The **challenge** rate Ki is a constant .

OK BAD OK OK OK OK OK OK

- 人工的に誤りを発生させて作成した学習データの有効性を調べる
- 低品質な学習データを作成し、モデルの精度の比較を行う
 - ・ 低品質なデータは、元のトークンに類似したトークンを挿入・置換することで作成

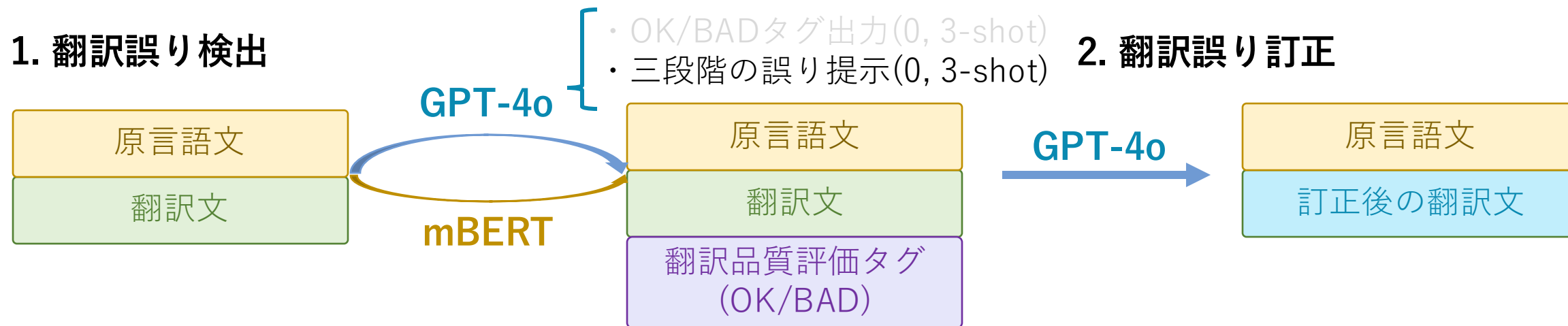
(a) 日英翻訳

モデル	項目	Precision	Recall	F1
未学習 mBERT	OK	0.673	0.140	0.232
	BAD	0.372	0.882	0.523
	TOTAL	F1: 0.338, MCC: 0.031		
低品質教師データ	OK	0.707	0.913	0.797
	BAD	0.694	0.342	0.459
	TOTAL	F1: 0.673, MCC: 0.320		
提案手法	OK	0.855	0.883	0.869
	BAD	0.785	0.740	0.762
	TOTAL	F1: 0.830 , MCC: 0.631		

(b) 英日翻訳

モデル	項目	Precision	Recall	F1
未学習 mBERT	OK	0.667	0.002	0.004
	BAD	0.362	0.998	0.531
	TOTAL	F1: 0.195, MCC: 0.002		
低品質教師データ	OK	0.747	0.873	0.805
	BAD	0.681	0.478	0.562
	TOTAL	F1: 0.717, MCC: 0.388		
提案手法	OK	0.870	0.880	0.875
	BAD	0.784	0.769	0.776
	TOTAL	F1: 0.839 , MCC: 0.651		

1. 翻訳誤り検出



2. 翻訳誤り訂正

● テストデータ

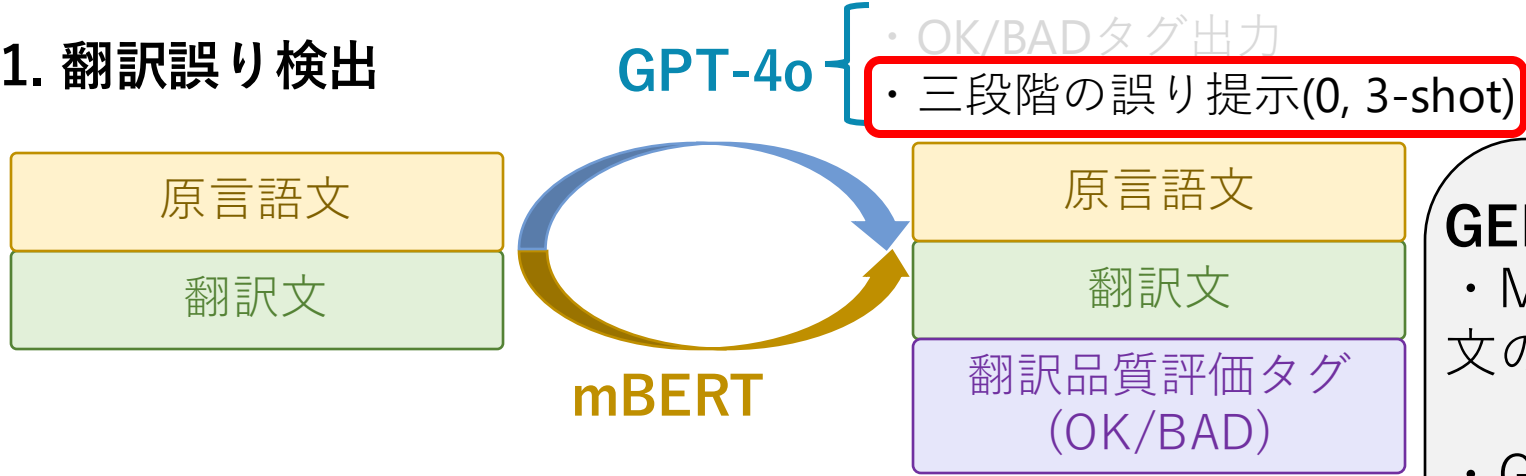
1. 人工誤り特許データ：200文
2. 繰り返し誤り特許請求項(翻訳文の文長が参照文の2倍以上)：211文
3. 訳抜け誤り特許請求項(翻訳文の文長が参照文の0.5倍以下)：200文

● 評価指標

・ BLEU

- ・ 翻訳文と参照訳文との一致度を評価
- ・ 特許のような用語の厳密な翻訳が必要な場合は、BLEUが重視される

1. 翻訳誤り検出



例 (GEMBA-MQM)

原言語文：なお、変化率 K_i は定数である。

目的言語文：The temperature change rate K_i of a sample .

GEMBA とは？ [kocmi+, WMT2023]

- ・ MQM形式(3段階のエラー形式)で文の翻訳品質評価を行う手法
- ・ GPT-4を用いて翻訳エラー範囲・種類を出力を行う
- ・ 論文中の評価ではGPT-4を用いた最も高い誤り検出精度を達成する手法

MQM annotations:

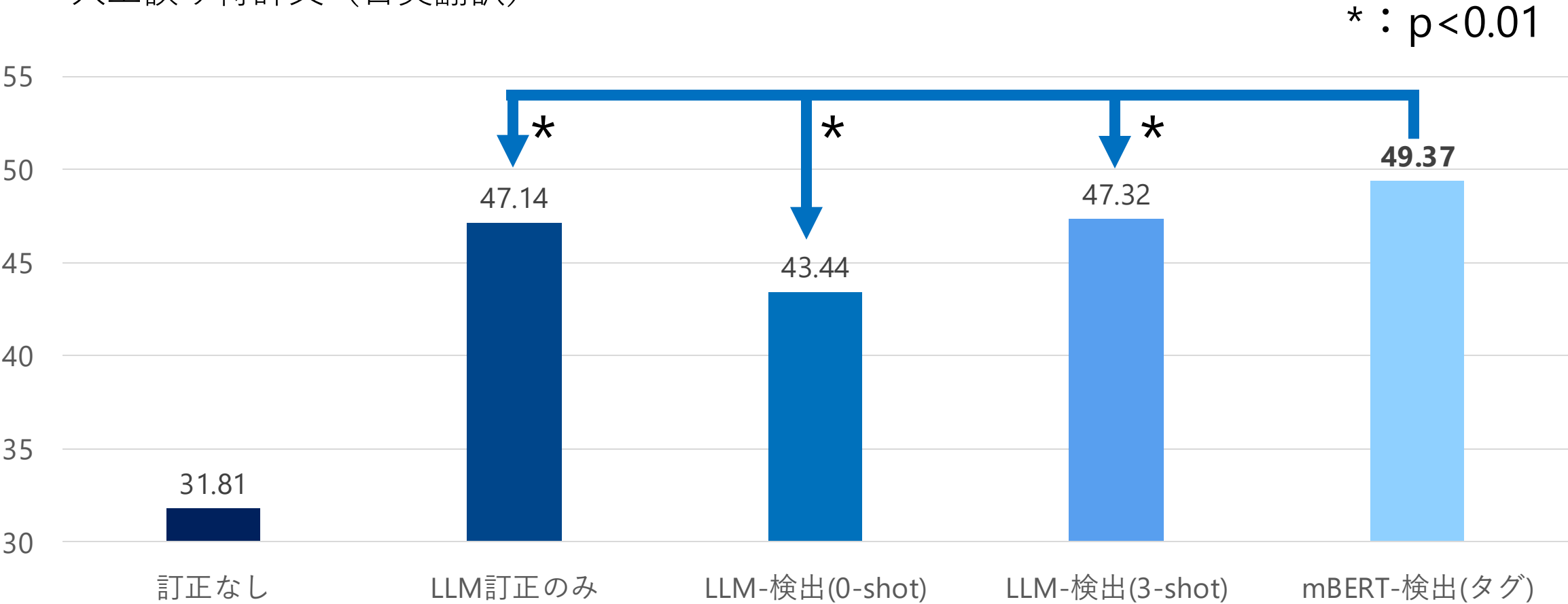
Critical:
accuracy/mistranslation - "The . . . sample" (The translation fails to convey that the original sentence states K_i 's change rate is constant.)

Major:
accuracy/omission - "なお" (The word "なお" is omitted)

Minor:
no-error

● 5 種類の手法の翻訳訂正評価を行う（BLEU）

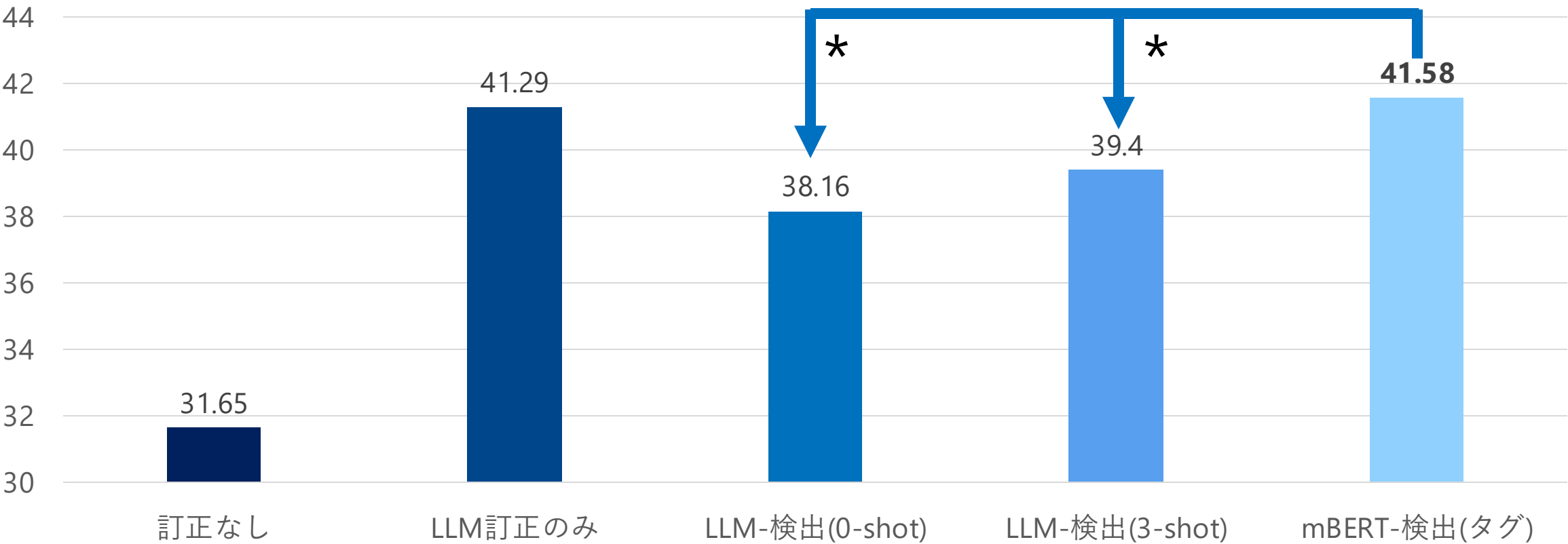
・ 人工誤り特許文（日英翻訳）



● 5 種類の手法の翻訳訂正評価を行う（BLEU）

・ 人工誤り特許文（英日翻訳）

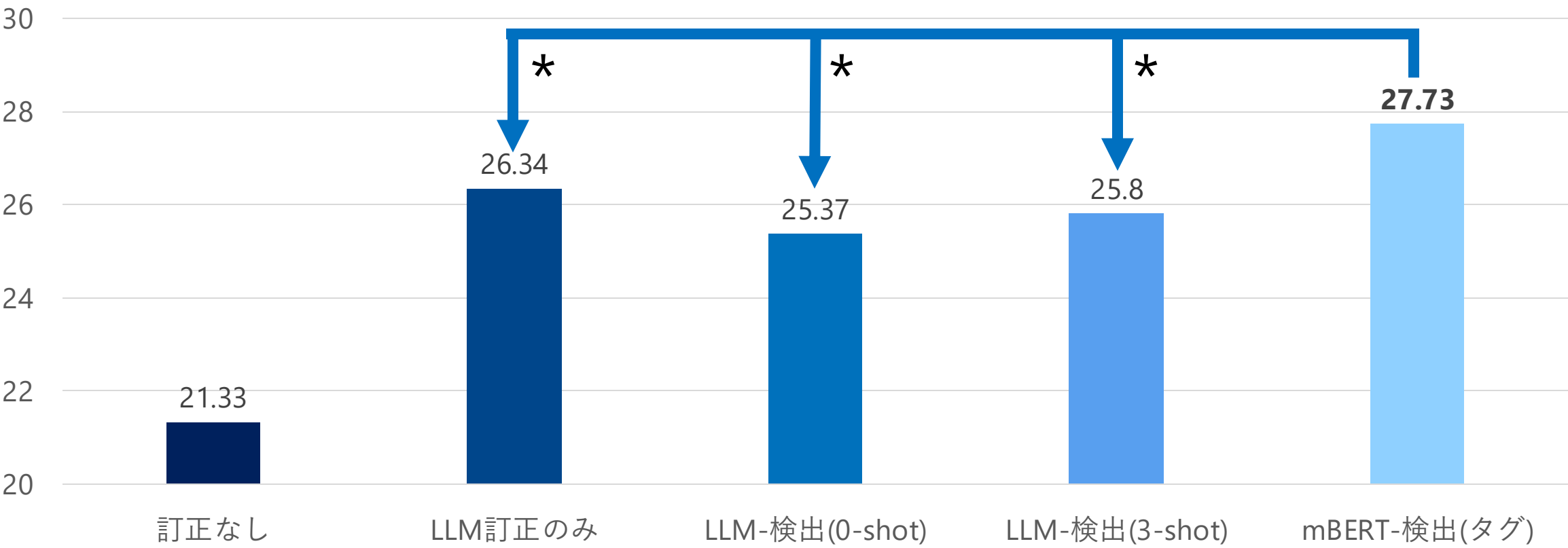
* : $p < 0.01$



● 5 種類の手法の翻訳訂正評価を行う（BLEU）

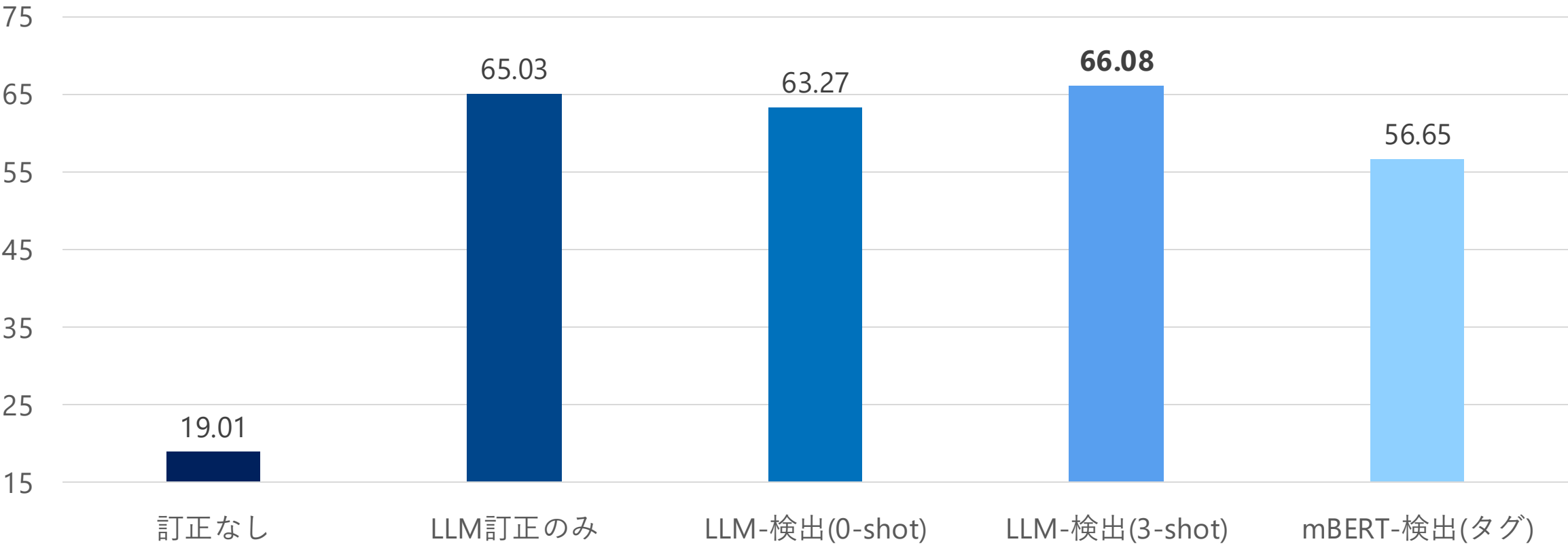
・ 繰り返し誤り特許請求項（日英翻訳）

* : $p < 0.01$



● 5 種類の手法の翻訳訂正評価を行う（BLEU）

・ 訳抜け誤り特許請求項（日英翻訳）



● 事前学習済みモデルとLLMを用いた2段階の手法を提案

1. 翻訳誤り検出：

- ・ 作成した特許文の擬似データで学習
 - ⇒ 人手データがなくても誤り検出モデルの訓練が可能に
- ・ encoderモデルを高品質な訓練データで訓練することで、他タスクで高い性能を示すLLMを上回る精度を達成

2. 翻訳誤り訂正：

- ・ mBERTでの翻訳誤り検出結果を用いる
 - ⇒ LLMでの翻訳誤り訂正精度が統計的に有意に向上
- ・ 翻訳誤り検出位置を修正 ⇒ 不必要な訂正は行わない

● 今後の展望

- OK/BADだけでなく、翻訳品質タグの改善
- 学習の特許の文量を増やす

Appendix

System Prompt

You are a translation checker.

You will be given:

- 1) A Japanese sentence (source text).
- 2) An English sentence (the current translation).
- 3) A list of token-level annotations (BAD/OK) for the English sentence.

Your tasks are:

- 1. Identify translation errors or inaccuracies in the English sentence relative to the Japanese source.
 - Use the BAD/OK annotation list as a reference, but also rely on your own judgment.
- 2. Propose corrections or improvements for each identified error.
- 3. Provide a final, corrected English translation that reflects all improvements.

Output Format:

[Translation Errors]

- (1) <具体的にどの部分が誤りか、どのように修正すべきか>
- (2) <...>
- ...

[Corrected Translation]

<最終的に修正を反映した正しい英文>

Constraints:

- Do not provide explanations or commentary beyond what is requested in the Output Format.
- Keep your output concise and organized.

User Prompt

Japanese source sentence:
{source_text.strip()}

English translation to check:
{translated_text.strip()}

Token-level annotation:
{annotation_list.strip()}

Please:

- 1. List errors and their corrections under [Translation Errors].
- 2. Provide the corrected translation under [Corrected Translation].

English source: I do apologise about this, we must gain permission from the account holder to discuss an order with another person, I apologise if this was done previously, however, I would not be able to discuss this with yourself without the account holders permission.

German translation: Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen, um eine Bestellung mit einer anderen Person zu besprechen. Ich entschuldige mich, falls dies zuvor geschehen wäre, aber ohne die Erlaubnis des Kontoinhabers wäre ich nicht in der Lage, dies mit dir involvement.

MQM annotations:

Critical:

no-error

Major:

accuracy/mistranslation - "involvement"

accuracy/omission - "the account holder"

Minor:

fluency/grammar - "wäre"

fluency/register - "dir"

English source: Talks have resumed in Vienna to try to revive the nuclear pact, with both sides trying to gauge the prospects of success after the latest exchanges in the stop-start negotiations.

Czech translation: Ve Vídni se ve Vídni obnovily rozhovory o oživení jaderného paktu, přičemž obě partaje se snaží posoudit vyhlídky na úspěch po posledních výměnách v jednáních.

MQM annotations:

Critical:

no-error

Major:

accuracy/addition - "ve Vídni"

accuracy/omission - "the stop-start"

Minor:

terminology/inappropriate for context - "partaje"

Chinese source: 大众点评乌鲁木齐家居商场频道为您提供高铁居然之家地址, 电话, 营业时间等最新商户信息, 找装修公司, 就上大众点评

English translation: Urumqi Home Furnishing Store Channel provides you with the latest business information such as the address, telephone number, business hours, etc., of high-speed rail, and find a decoration company, and go to the reviews.

MQM annotations:

Critical:

accuracy/addition - "of high-speed rail"

Major:

accuracy/mistranslation - "go to the reviews"

Minor:

style/awkward - "etc.,"

1 つ目の例

・ 英語 – ドイツ語

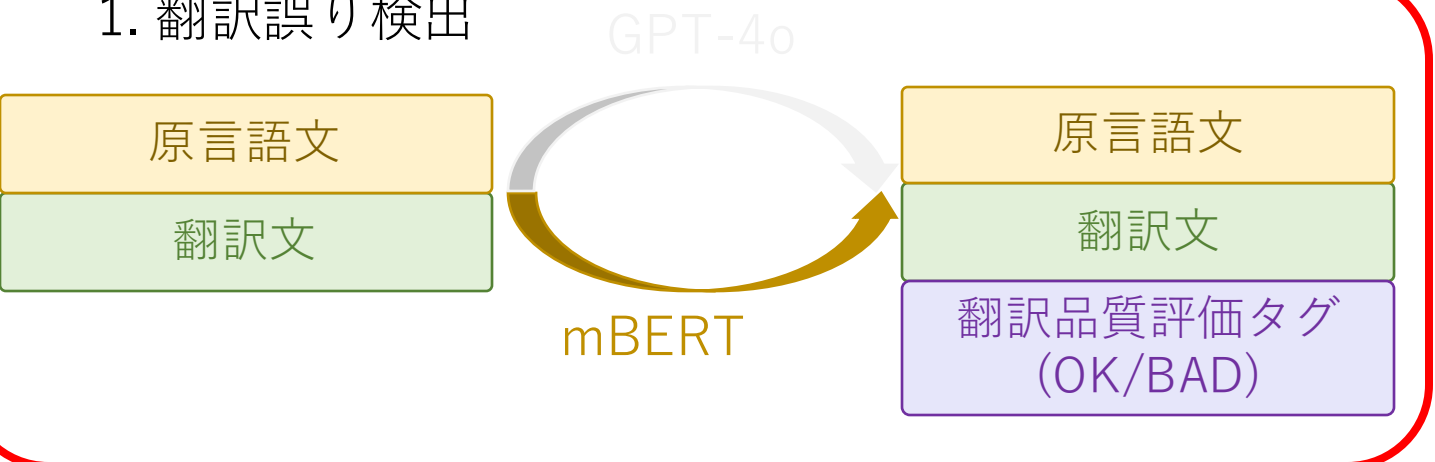
2 つ目の例

・ 英語 – チェコ語

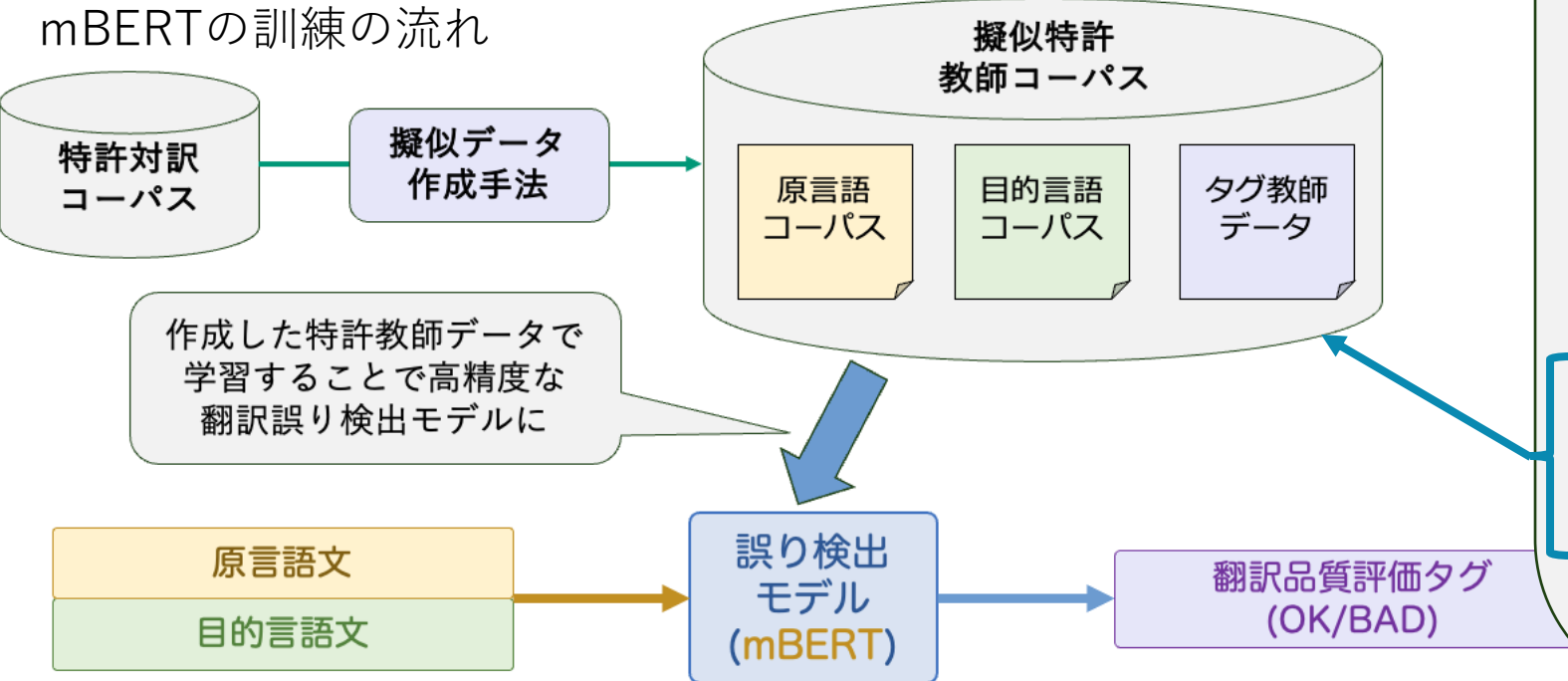
3 つ目の例

・ 中国語 – 英語

1. 翻訳誤り検出



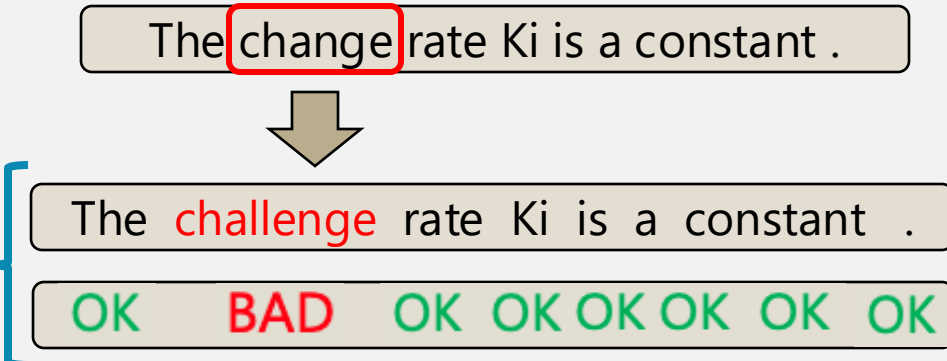
mBERTの訓練の流れ



擬似データ作成手法

- ・ 対訳特許文に以下の手法で人工的な誤りを発生させる
- 1. トークンを5%の確率で削除
- 2. トークンを10%の確率で挿入
- 3. トークンを30%の確率で置換

例 (置換した場合)



原文

前記 NK 細胞は、血液または細胞株に由来し、好ましくは、細胞株に由来し、より好ましくは、前記細胞株に由来する NK は NK92 細胞株であることを特徴とする 請求項 12 に記載の免疫細胞。

参照訳文

The immune cell of claim 12, wherein the NK cell is derived from blood or a cell line; preferably, from a cell line; and more preferably, the NK cell from a cell line is NK92 cell line.

MT 文

The immune cell of claim 12, wherein the NK cell is derived from blood or a cell line.

提案手法

The immune cell of claim 12, characterized in that the NK cell is derived from blood or a cell line, preferably from a cell line, and more preferably from the aforementioned cell line, specifically the NK92 cell line.

- ・ F1値(0~1)

$$F1 = \frac{2 \times (Precision \times Recall)}{Precision + Recall}$$

- ・ MCC(-1~1)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$