日本語の**古語**
（本研究では平安時代末期
～江戸時代）

古典日本語の現代語機械翻訳のための評価資源の整備

東山 翔平[1]　大内 啓樹[2]　橋本 雄太[3]　藤田 篤[1]

[1]情報通信研究機構　[2]奈良先端科学技術大学院大学　[3]国立歴史民俗博物館

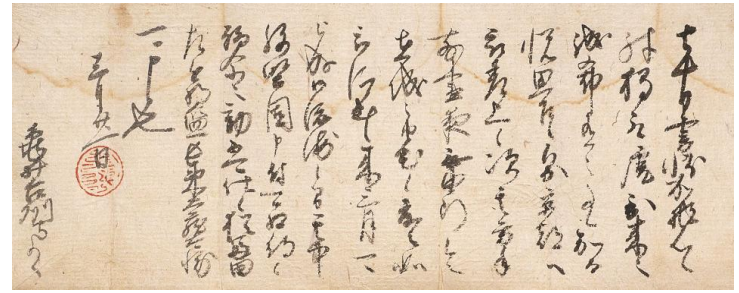
本発表は、以下の発表内容に基づくものです：
- 東山 他，「中世・近世日本語資料の現代語機械翻訳における自動評価指標の検証」，じんもんこん2025.
- 東山 他，「中世・近世日本語資料の現代語機械翻訳：評価用対訳データセットの構築とLLMの性能評価」，NLP2026.

歴史的資料の読解支援に向けて

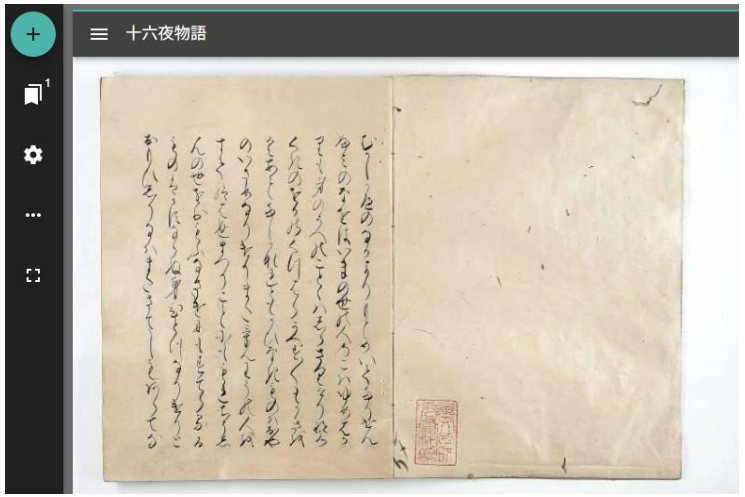
●背景

- 日本の歴史的な文字資料は、膨大な数が遺っている
  - 江戸時代の古文書（こもんじょ）・古記録20億点[1]／和本数百万点[2]など存在
- 約30年来でデジタルアーカイブの整備が進み、数十万点の資料画像にデジタルアクセス可能に
- しかし、テキスト情報／書かれた内容は、（一般の人／計算機にとって）未活用といえる

近世古文書の例：「豊臣秀吉朱印状」
（原画像出典： https://chuseimonjo.net/#/document/36）



国書データベース 収録資料の例：
「十六夜物語」(国文学研究資料館所蔵)



●本研究プロジェクトの目標

一般の人でも歴史的資料を読み解き、過去の知識にアクセス可能とする「古語の壁の解消」 ＝究極的ゴール

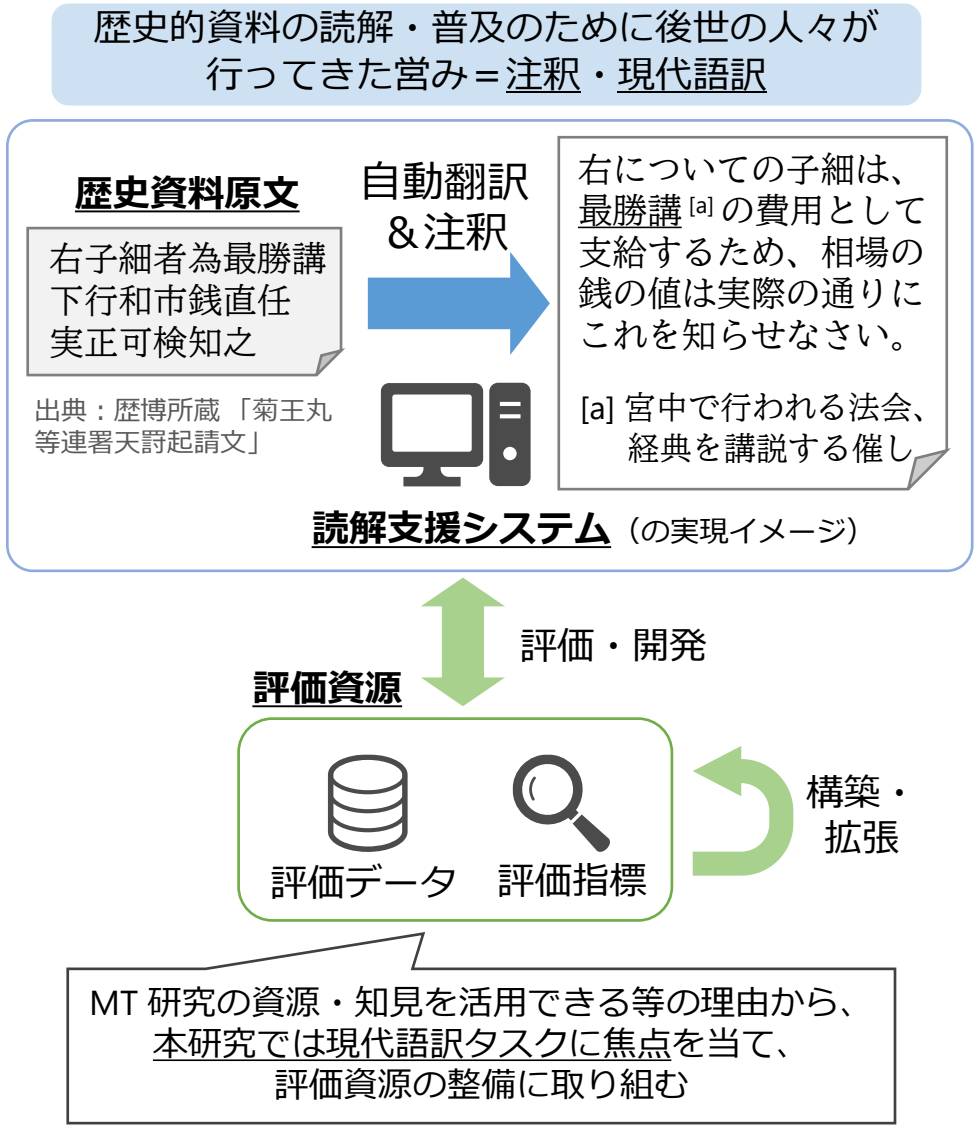
のための読解支援技術の発展、研究の推進

[1] 奥村弘『なぜ地域歴史資料学を提起するのか』より
[2] 中野三敏『和本のすすめ』より

読解支援としての現代語機械翻訳

●課題：評価資源の不足
- システムの評価に使える評価データがない
- 自動評価指標の有効性が未知

●本研究の貢献：評価資源の整備＋評価
1. 古語・現代語の評価用対訳データセット “JHPT” を構築・公開
2. 既存の MT 自動評価指標をメタ評価し、古語・現代語 MT での有効性を検証・確認
3. 既存 LLM の現代語訳精度を評価し、モデル／ドメイン間の精度差に関する知見を明らかにした



データセット公開URL：https://github.com/nict-astrec-att/jhpt

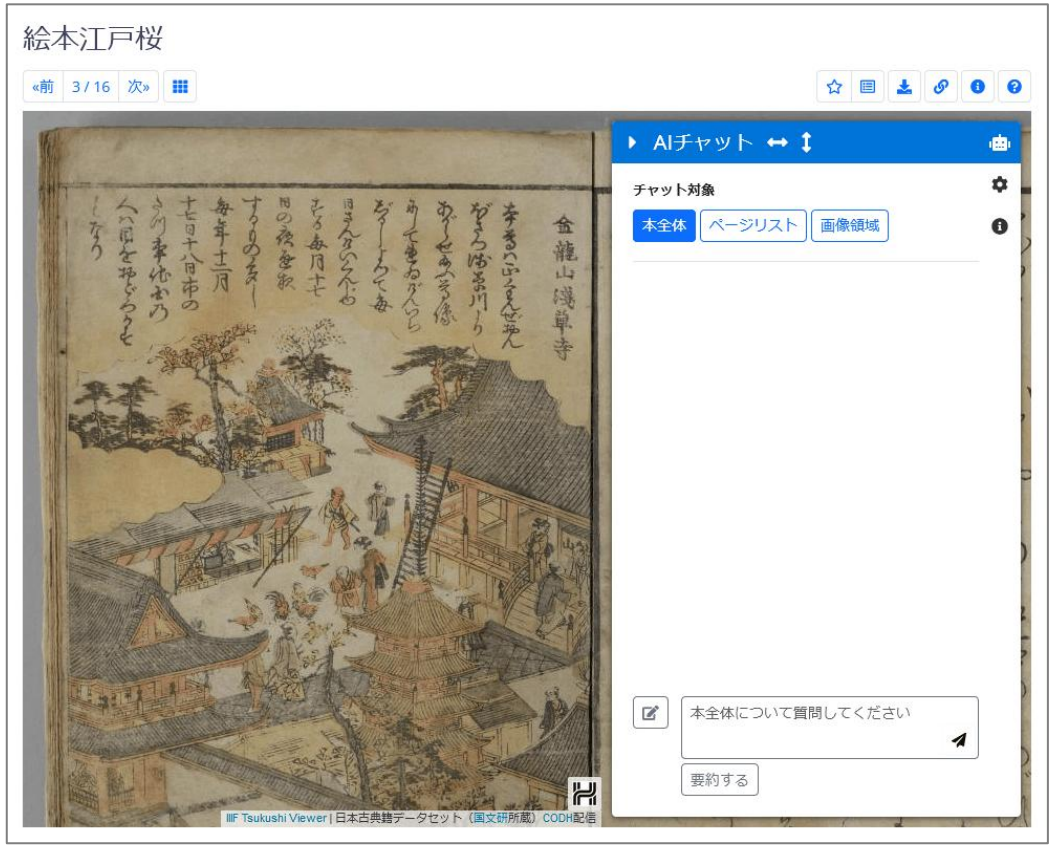
関連研究

関連研究：日本の文学/歴史的資料の読解支援

- 商用 LLMを利用した対話アプリ
  - ユーザ有用性に主眼が置かれた研究。定量的評価を焦点とする本研究とは直交・補完的

HUMANITEXT
AOZORA

https://aozora.humanitext.ai/



IIIF Tsukushi Viewer [北本+ '24]

https://codh.rois.ac.jp/software/iiif-tsukushi-viewer/

関連研究：古語→現代語の機械翻訳

- ラテン語 [Volk+ ‘24]
  - 商用 NMT と商用 LLM（GPT-4）の翻訳精度を評価

- 仏教漢語（Buddhist Chinese） [Nehrdich+ ‘23; ‘25]
  - 対訳データ（21万文対）の構築、NMT/LLM の微調整・翻訳精度評価、
    自動評価指標のメタ評価、などを実施

- 日本の古典文学 [星野+ ‘15; Usui+ ‘23]
  - 非公開コーパス（小学館『新編日本古典文学全集』）から作成された対訳を使用し、
    SMT や NMT を学習・評価
    - ➢ ただし、後続の研究者が同コーパスを利用して研究を行うことや、
      成果物を一般公開することは困難

関連研究：古語→現代語の機械翻訳

- 日本の古文書（こもんじょ）[1] [橋本+ ‘25]
  - 中世・近世古文書を商用 LLM に現代語訳させ、品質を人手評価（文単位および単語単位）
  - 人手評価の過程で、既存の／新規に人手作成した現代語訳を利用

橋本ら公開のオリジナルデータ：現代語訳・LLM生成訳・人手評価結果（図では単語単位）

区分		単語									正答数	単語等数
原文		如	本	可	為	寺領	之	旨	也、	者、		
訳		もとのように寺領であるべきとのことである、ということだ。										
Claude	生成訳	以前の通り寺領とすべき旨を示されたい。										
	点数	1	1	1	1	1	1	1	意訳	1	8	9
Gemini	生成訳	本のように寺の領地とすべきである。										
	点数	1	0	1	1	1	0	0	1	1	6	9
ChatGPT	生成訳	それが寺の領地としての本来の意図です。										
	点数		0	0	0	1	0	0	1	1	3	9
DeepSeek	生成訳	元の通り寺領とするよう命じるべきである。										
	点数	1	1	0	1	1	0	0	1	1	6	9

本研究にて活用

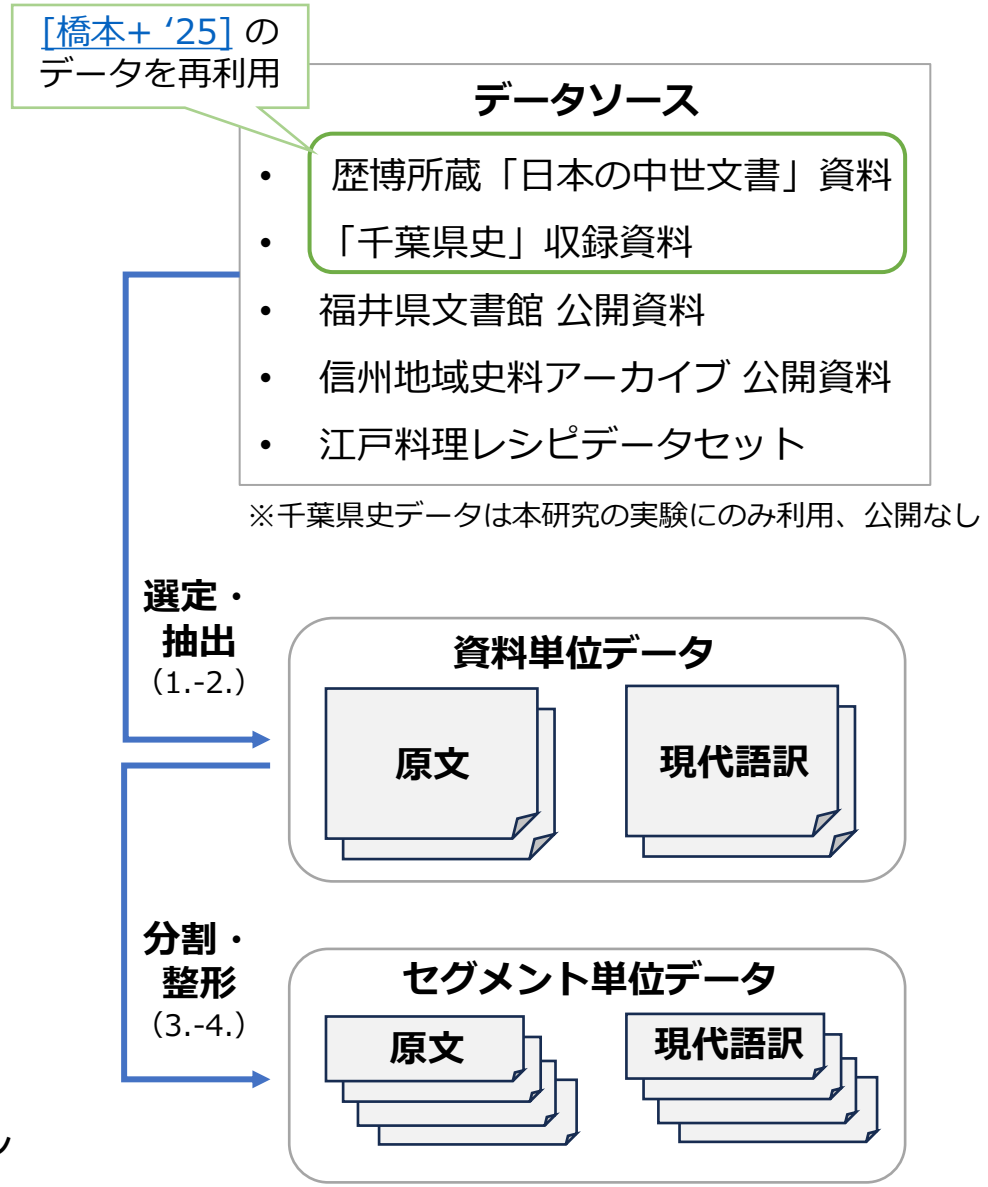
- 原文・現代語訳を対訳データの一部に利用
- LLM 生成訳とともに、0～1に換算した人手評価スコアを、メタ評価に利用

[1] 「古文書」は、歴史学研究の素材となる文献史料のうち、差出者から受取者へ宛てて書かれたもの（公的／私的な書状）。差出者・受取者のない文献史料は「古記録」にあたる。

貢献①：対訳データセット構築

対訳データセットの構築フロー

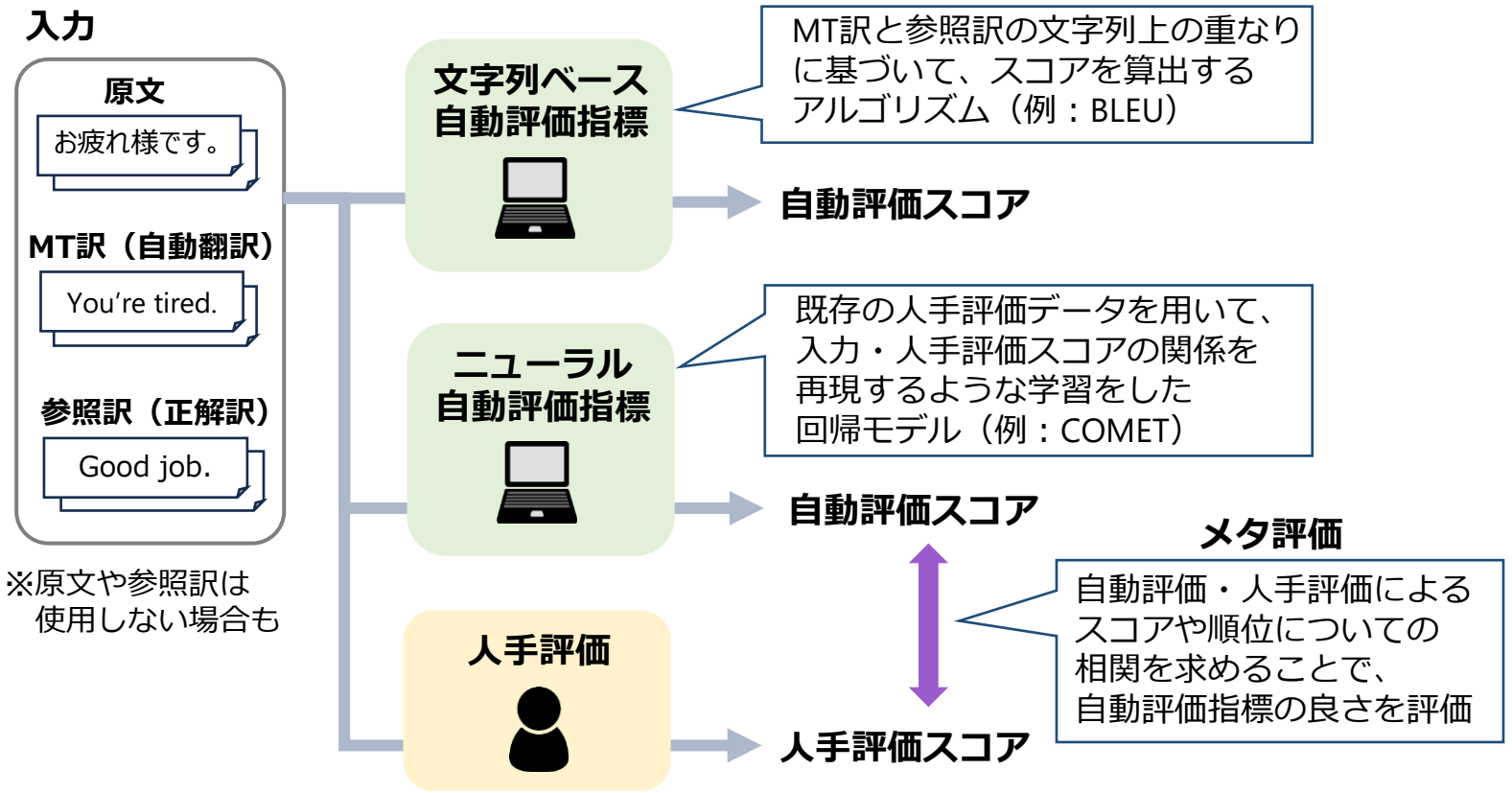
1. データソース・資料の選定
   - 原文、現代語訳とも入手・再公開可能なものを選択
2. テキスト抽出
   - HTML/PDF ファイル等から必要なテキストを取得
3. セグメント分割・対訳対応付け
   - 「単独で内容を理解できる意味的なまとまり」（読点や句点までの位置）をセグメントと認定
   - 各原文セグメントに訳文セグメントを対応付け
4. テキスト整形・対象外フラグ付与
   - 原文／現代語訳テキストに含まれるルビや注釈を分離
   - 人名・日付のみなど翻訳不要のセグメントにフラグ付与し評価対象から除外



貢献②：機械翻訳自動評価指標の有効性検証

自動評価指標と、その検証の必要性

- 多言語に対応したニューラル指標が提案され、メタ評価（＝人手評価結果との相関の計算）の結果、文字列ベース指標よりも良い（＝相関が高い）と報告されている
- 検証結果の汎化性には限界がある。特に、未検証の翻訳方向については新たに検証が必要

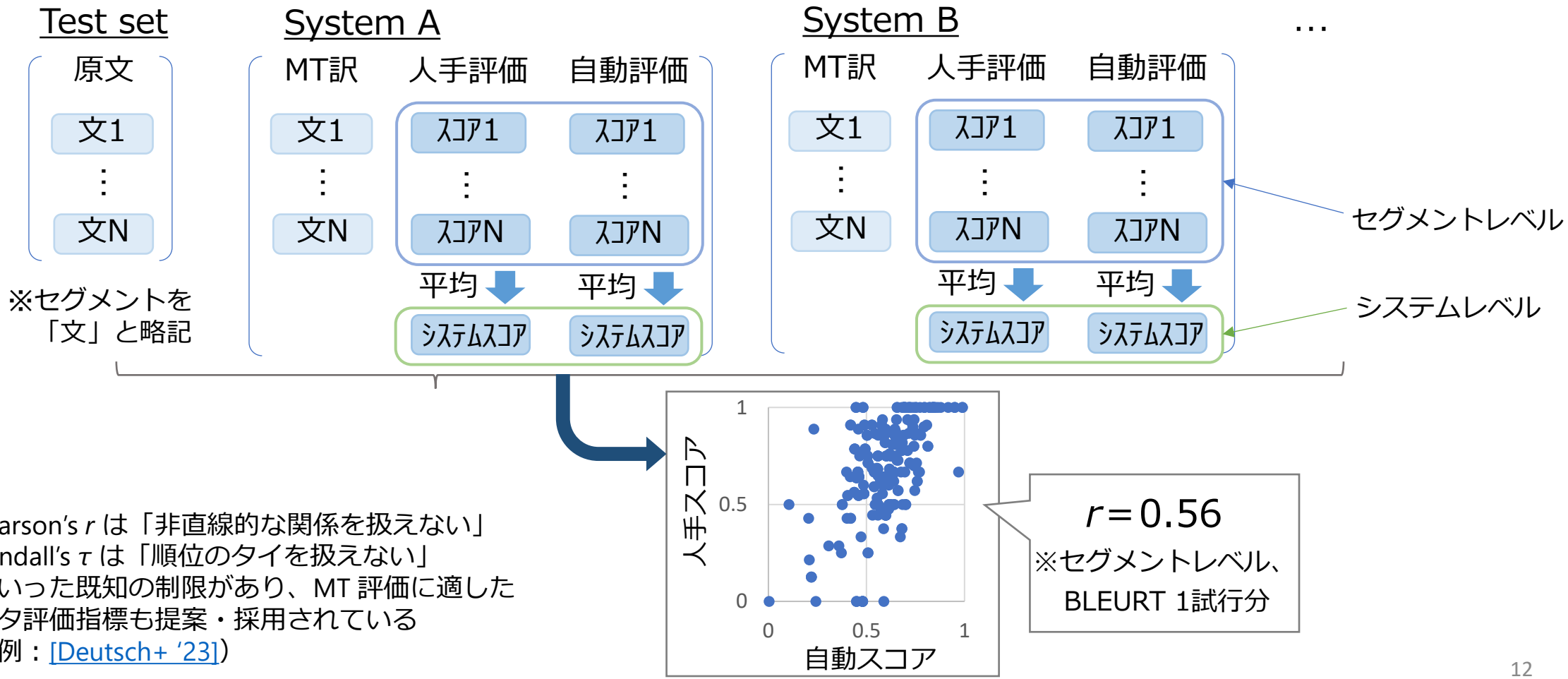


（参考）COMET-22 についての System-level Pearson correlation [Freitag+ '22]

en-de	0.771	専門家 MQM 評価スコアとの相関
en-ru	0.900	
zh-en	0.942	
cs-en	0.821	クラウドワーカー Direct Assesment スコアとの相関
de-en	0.446	
ja-en	0.976	
ru-en	0.857	
uk-en	0.714	
zh-en	0.135	

メタ評価の概要（Pearson の積率相関係数 r の場合）

– 各セグメント/システムの人手・自動スコア対をデータ点とし、
人手スコア列と自動スコア列の間の相関係数を算出する



※Pearson's r は「非直線的な関係を扱えない」
Kendall's τ は「順位のタイを扱えない」
といった既知の制限があり、MT 評価に適した
メタ評価指標も提案・採用されている
（例：[Deutsch+ '23]）

実験：LLM 4モデルの自動評価結果のメタ評価

- 設定
  - [橋本+ ‘25] のデータを再利用し、中世・近世古文書資料の原文67セグメントに対する4システム（Claude Sonnet 3.5, Gemini 1.5, GPT-4o, DeepSeek-R1）の生成訳を利用
  - 「4システム×67セグメント」(Segment level) 、「4システム」(System level) について、自動評価（6指標のいずれか）と人手評価のスコア/順位間の相関を求める
  - 67セグメントから復元抽出する（各試行でN事例から重複を許してN事例を抽出）Bootstrap resampling を1000試行行い、相関係数値の平均と95%信頼区間を求めた

メタ評価に用いた6指標

タイプ	評価指標	ツール/モデル	学習方法	日本語のMT人手評価データでの微調整	入力
文字列	BLEU	SacreBLEU	なし	-	(MT訳, 参照訳)
文字列	chrF	SacreBLEU	なし	-	(MT訳, 参照訳)
深層学習	BERTScore	tohoku-nlp/bert-base-japanese-v3	事前学習	-	(MT訳, 参照訳)
深層学習	BLEURT	BLEURT-20	微調整	なし（他の多言語データ）	(MT訳, 参照訳)
深層学習	COMET	wmt22-comet-da	微調整	日⇔英 人手評価データ使用	(原文, MT訳, 参照訳)
深層学習	CometKiwi	wmt22-cometkiwi-da	微調整	日⇔英 人手評価データ使用	(原文, MT訳)

実験：LLM 4モデルの自動評価結果のメタ評価

評価指標	Pearson's r	Kendall's τ
Segment level		
BLEU	0.441 [0.299, 0.507]	0.290 [0.191, 0.383]
chrF	0.462 [0.356, 0.563]	0.320 [0.213, 0.427]
BERTScore	0.489 [0.381, 0.591]	0.328 [0.221, 0.440]
BLEURT	**0.554** [**0.424**, **0.656**]	**0.378** [**0.269**, **0.480**]
COMET	0.516 [0.390, 0.627]	0.353 [0.257, 0.443]
CometKiwi	0.222 [0.073, 0.358]	0.122 [0.012, 0.226]

評価指標	Pearson's r	Kendall's τ
System level		
BLEU	0.897 [0.655, 0.996]	0.843 [0.333, **1.000**]
chrF	**0.937** [**0.779**, **0.999**]	**0.917** [**0.667**, **1.000**]
BERTScore	0.885 [0.708, 0.986]	0.876 [**0.667**, **1.000**]
BLEURT	0.855 [0.655, 0.975]	0.845 [0.333, **1.000**]
COMET	0.814 [0.543, 0.974]	0.766 [0.333, **1.000**]
CometKiwi	-0.045 [-0.714, 0.653]	0.020 [-0.333, 0.667]

- 結果
  - CometKiwi を除き、セグメントレベルでは中程度の相関、システムレベルでは強い相関
    - 評価指標によるスコアの勝敗によって、「2セグメント間の優劣を一部判定可能」、「2システム間の優劣を判定可能」、と解釈できる
    - 5指標について、**一定の有効性を確認**
  - ただし、少ないセグメント数（=67）、僅かなシステム数（=4）での評価のため、セグメント・システム数を増やしても同様の傾向となるか、より信頼性の高い結論を得るために追加の検証も必要（今後実施予定）

貢献③：既存 LLM の翻訳精度自動評価（+定性的分析）

実験：LLM の翻訳精度自動評価

メタ評価に使用した
67セグメントも含む

●設定
- データ：構築した対訳データセット全体（726セグメント）を評価に使用
- 指標：メタ評価から、自動評価指標は BLEU、BLEURT を使用
- モデル：商用 LLM 6モデル、オープン LLM 6モデルを評価

6モデル以外にも、Qwenシリーズ、gpt-oss-20b、
Karamaru なども一部評価したが、
顕著な結果は見られなかったため割愛

●推論・評価方法
- 忠実な現代日本語訳を求める
  右のプロンプトで Zero-shot 推論
  （微調整なし）
- データセットを5ドメインに分けつつ、
  全ドメイン／ドメイン別でそれぞれ評価

以下は日本の歴史資料の文章です。原文を現代日本語に翻訳した現代語訳を出力し、他には何も出力しないでください。現代日本語として自然な文字・語彙・表現・文法を用いて、原文の内容に忠実な訳としてください。英単語の使用を禁止します。
原文：{source_text}
現代語訳：

実験：LLM の翻訳精度自動評価（5ドメイン全体）

– 商用 LLM、特に Claude が高精度
  - Gemini/GPT で思考レベルを上げると精度微増

参考：メタ評価対象の67セグメントに対する Claude Sonnet 3.5 の評価結果 [東山 '25]

BLEU	BLEURT	人手評価（文単位）	人手評価（単語単位）
24.7	66.6	62.7%	81.1%

から、品質は改善の余地ありと想定される。

– オープンモデルの中では、Gemma-2-Llama-Swallow や DeepSeek は健闘し、GPT-5/5.2 に近いスコア

ドメイン別スコア（次頁）のマクロ平均

LLM	全データ BLEU	全データ BLEURT
Claude Opus 4.5	**28.5**	**66.6**
Claude Sonnet 4.5	26.4	65.5
Gemini 3 Flash (thinking_level: medium)	21.3	**66.5**
Gemini 3 Flash (thinking_level: minimal)	20.9	65.8
Gemini 2.5 Flash (thinking_budget: 0)	23.9	64.8
GPT-5.2 (reasoning effort: medium)	22.6	64.3
GPT-5.2 (reasoning effort: none)	21.6	63.2
GPT-5 (reasoning effort: minimal)	19.7	62.9
DeepSeek-R1 (thinking: off)	19.7	**63.8**
Gemma-2-Llama-Swallow-27b-it-v0.1	**21.2**	62.2
Gemma-2-Llama-Swallow-9b-it-v0.1	19.0	61.2
llm-jp-3.1-13b-instruct4	19.4	57.9
Llama-3-ELYZA-JP-8B	17.5	54.9
Sarashina2.2-3b-instruct-v0.1	15.1	61.3

実験：LLM の翻訳精度自動評価（ドメイン別）

– 各 LLM とも「古文書・古記録」ドメインで精度が低い傾向
  - ➢ 存在する資料自体が少なく、当該ドメインテキストでの学習があまりされていない
  - ➢ 漢文形式のテキストが多く、現代日本語と乖離が大きいことから出力品質低下した

} と想定

LLM	中世古文書・古記録		近世古文書・古記録		近世古典籍（文学・地誌）		近世古典籍（記録等）		近世古典籍（料理）	
	BLEU	BLEURT	BLEU	BLEURT	BLEU	BLEURT	BLEU	BLEURT	BLEU	BLEURT
Claude Opus 4.5	**27.2**	**65.8**	**23.0**	62.8	**38.4**	**70.4**	**27.5**	68.4	**26.5**	**65.4**
Claude Sonnet 4.5	23.6	65.2	22.0	61.7	36.9	69.0	25.4	**68.6**	24.3	63.2
Gemini 3 Flash (thinking_level: medium)	19.4	64.3	19.3	**64.6**	30.1	70.0	17.9	68.5	19.8	65.2
Gemini 3 Flash (thinking_level: minimal)	18.8	63.6	19.2	64.3	29.5	69.0	17.4	67.0	19.4	65.1
Gemini 2.5 Flash (thinking_budget: 0)	21.2	62.8	20.2	61.1	33.0	67.7	21.6	67.6	23.5	64.7
GPT-5.2 (reasoning effort: medium)	18.6	61.3	19.5	61.2	29.6	68.0	22.3	66.8	23.1	64.0
…										
DeepSeek-R1 (thinking: off)	15.6	60.6	16.6	**60.8**	26.4	**66.6**	18.3	66.6	21.4	**64.3**
Gemma-2-Llama-Swallow-27b-it-v0.1	**19.5**	**60.7**	16.9	59.2	**26.6**	65.5	20.6	**67.1**	**22.3**	58.6
Gemma-2-Llama-Swallow-9b-it-v0.1	16.0	57.3	15.6	58.4	24.3	63.4	19.2	65.8	20.1	61.3
llm-jp-3.1-13b-instruct4	17.6	52.3	15.5	56.0	26.9	62.0	17.4	62.3	19.4	56.9
…										

※全モデルの結果はNLP2026論文を参照

翻訳結果事例1（千葉県史 近世古文書資料）

– Claude／Gemini は謝罪の中心的な文意を伝えているが、他2モデルでは失敗

御番所を通り過ぎた　不始末を謝罪

原文	此段前文申上候通り楫痛殊ニ汐行悪敷候故、是迄乗参御番所乗越候段不調法可申上様無御座候、
参照訳	これは前に申し上げた通り楫が故障して、そのうえ潮流が悪かったので今までは寄港していた浦賀番所を素通りしてしまったもので、不始末で申し訳ありません。
Claude Opus 4.5	この件は前文で申し上げました通り、舵が痛んでおり、特に潮の流れが悪かったため、これまで乗ってきて御番所を乗り越えてしまいましたことは不調法であり、申し上げようもございません。
Gemini 3 Flash (thinking_level: medium)	この件については、以前申し上げました通り、舵の傷みに加えて、とりわけ潮の流れが悪かったため、ここまで航行してくる中で [Addition] [Omission: 乗参]番所を通り過ぎてしまったことは、不手際であり、弁解の余地もございません。
DeepSeek-R1 (thinking: off)	このたびは、先に申し上げましたように、櫓の痛みが特にひどく、潮の流れも悪いため、これまで通りに船を出して [Addition] [Omission: 乗参]番所まで参上することができず、まことに不都合でございますが、申し上げるべき方法もございません。
Gemma-2-Llama-Swallow-27b	この段の前文に申し上げました通り、楫の痛みと潮の流れが悪いため、これまで御番所に乗参して [Dependency]乗り越えてきたことは、不都合であると申し上げても差し支えございません。

赤：誤訳，青：その他誤り（Addition=過剰生成／Omission=訳し漏れ／Dependency=係り受け構造誤り）

翻訳結果事例2（料理）

エラー傾向の印象：モデルにとって未知であろう表現に対し、
その字面や文脈からそれっぽい表現を当てはめて、結果誤っている

- Claude ／ Gemini は、「よばして」を含む調理工程全体をほぼ適切に訳している。
  他2モデルは、「よばして」や他の個所で誤訳あり（下記以外のモデルも軒並み同様）。

この文脈では「ふやかす」の意

原文	是は 大麥を よく よばして ずいぶん 水にて能あらひ 小茶碗に すこし入レ たまごに昆布のだしに すこし 醤油と やき塩とを入レ よきかげんして 右茶碗にかけて むすべし
参照訳	これは、大麦をよくふやかし、水で丁寧に洗い、小茶碗に少し入れる。卵に昆布出汁と少量の 醤油、焼塩を入れ、丁度よい味加減にし、右の茶碗にかけて、蒸す。
Claude Opus 4.5	これは大麦をよく水に浸して [Addition]十分にふやかし、水でよく洗い、小さな茶碗に少し入れ、卵に昆布の出汁と少しの醤油と焼き塩を入れ、よい加減に調味して、その茶碗にかけて蒸すこと。
Gemini 3 Flash (thinking_level: medium)	これは、大麦をよくふやかして、水で十分に洗い、小さな茶碗に少し入れます。卵に昆布のだし、少々の醤油と焼き塩を加えて味を調え、それを先ほどの茶碗に注いで蒸してください。
DeepSeek-R1 (thinking: off)	これは、大麦をよく洗って、十分に水でよくすすぎ、小さい茶碗に少し入れ、卵と昆布のだし汁に少し醤油と焼き塩を加えて適当な味加減にし、先ほどの茶碗にかけて蒸らすのである。
Gemma-2-Llama-Swallow-27b	これは、大麦を良く洗い、たっぷりの [Addition]水でよく洗い、小さな茶碗に少し入れ、卵に昆布のだし汁に少し醤油と焼き塩を加えて、良い加減に煮て、その茶碗にかけ、混ぜ合わせます。

赤：誤訳，青：その他誤り（Addition=過剰生成）

まとめと今後の展望

●本研究の貢献

- 中世・近世資料を用いた評価用対訳データセットを構築・公開
- 既存 MT 評価指標について、メタ評価により一定の有効性を確認
- 既存 LLM の評価により、翻訳精度のモデル／ドメイン別傾向を確認

を通じて、「古典日本語の現代語機械翻訳」の研究分野推進のための研究に取り組んだ

●今後の展望

- 収録資料の時代・ドメイン・数量の点で、対訳データを拡大
- 低資源の状況での、オープン LLM の翻訳精度向上のための学習方法の検討
- コスト・信頼性・再利用性のバランスが良い人手評価方法の検討と、自動評価指標のメタ評価や自動エラー検出等への人手評価データの活用
- 英語や他の言語への翻訳方向拡大（に向けた評価資源の整備）を模索

付録：対訳データセット JHPT の詳細

●データサイズ（「千葉県史」データに由来する非公開分を除外）

- 詳細：https://github.com/nict-astrec-att/jhpt

データソース	#文書数	#セグメント数	#原文文字数	#参照訳文字数
歴博中世文書	12	84	1,828	2,846
福井県文書	7	29	776	1,062
信州地域史料	10	374	17,199	20,515
江戸料理レシピ	1	56	2,941	2,797
合計	30	543	22,744	27,220

●収録資料一覧

- https://github.com/nict-astrec-att/jhpt/blob/main/data02/doc_list.tsv