# 日本語検索タスクにおける機械翻訳テストコレクションの妥当性検証

岩間　悠莉† 　加藤　誠††,†††

† 筑波大学 知識情報・図書館学類　〒 305-–8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305-–8550 茨城県つくば市春日 1-2
††† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †yiwama@klis.tsukuba.ac.jp, ††mpkato@acm.org

**あらまし**　本研究では，機械翻訳テストコレクションと，人手で構築されたテストコレクションとを比較して，情報検索システムの性能評価においてどの程度一貫した評価を与えるのか明らかにすることを目的とする．本検証に向けて，英語・日本語間で同一内容であるとみなせるテストコレクションを構築した．このうち英語のテストコレクションを日本語に機械翻訳することで，構築方法のみが異なる同一言語・同一内容の 2 つのテストコレクションを得た．これらを用いて，複数の検索モデルによる検索評価を行い，評価結果の類似性を分析するとともに，異なる翻訳モデルを用いた場合の評価結果についても確認した．実験の結果，機械翻訳テストコレクションの一定の妥当性が示唆された．
**キーワード**　情報検索，評価・データセット，機械翻訳

## 1　はじめに

情報検索システムの性能を客観的に評価するためには，クエリ・文書・適合性判定から構成される評価用データセットであるテストコレクションが不可欠である．しかし，その整備状況には大きな言語間格差があり，英語については多様なテストコレクションが蓄積されている一方で，多くの英語以外の言語では十分なリソースが存在しない．さらに，高品質なテストコレクションの作成には人手による適合性判定が必要であり，そのコストは極めて高い．例えば，18 言語で構築された多言語テストコレクションでは，その構築に 5 人年を要していたと報告されている [22]．このような状況下で，英語以外の言語における情報検索研究や実システムの評価は，英語と比較して利用可能なテストコレクションの制約を受けている．

英語以外の言語でのテストコレクション不足を補う手段として，既存のテストコレクションを機械翻訳して利用する試みが既に行われている．しかし，これらの研究においてはその評価の妥当性が検証されていないことがある [11], [20]．あるいは翻訳前のテストコレクションと翻訳後の他言語テストコレクションとの評価結果を比較し，両者が近いほど望ましいと暗黙に仮定している [6]．

ここで重要となるのは，翻訳後の言語において人手で構築された同一内容のテストコレクションが存在した場合に，機械翻訳テストコレクションが，情報検索システムの性能評価において同様の結論を与えられるかという点である．人手で構築されたテストコレクションは，対象言語における情報検索評価の基準として広く用いられており，一般に実際の情報要求を反映していると考えられている．機械翻訳テストコレクションの妥当性については，このような人手テストコレクションとの比較を通じて，評価結果がどの程度整合するかという観点から検討することが，1 つの有効なアプローチである．ここで「同一内容」とは，クエリが同一の情報要求を表し，文書および適合性判定についても同一の情報を伝えているとみなせる設定を意味する．また，本研究において「同様の評価」とは，評価値の絶対的な一致を直接比較するのではなく，検索モデル間の相対的な性能比較やモデル順位，文書ランキングの傾向といった評価結果の構造に着目した比較を指す．しかし，既存研究ではこのような観点からの直接的な比較は行われておらず，機械翻訳テストコレクションによる評価の妥当性は十分に検証されていない．

本研究の目的は，機械翻訳テストコレクションと，人手で構築された同一言語・同一内容のテストコレクションと比較して，情報検索システムの性能評価においてどの程度一貫した結論を与えるのかを定量的に明らかにすることである．そのために本研究では，英語と日本語の 2 言語に焦点を当て，同一内容を表すクエリおよび文書が言語間で対応付けられたクエリ集合・文書集合を組み合わせることで，テストコレクションを構築する．次に，このうち英語のクエリと文書を日本語へ機械翻訳することで「人手で作成された日本語テストコレクション」と「英語から機械翻訳された日本語テストコレクション」という 2 種類の日本語テストコレクションが得られる．これらは構築方法のみが異なる，同一言語・同一内容のテストコレクションとみなせる．これを用いて，複数の検索モデルによる評価実験を行い，人手テストコレクションと機械翻訳テストコレクションそれぞれに対して得られる評価値や検索モデルの順位を比較することで，機械翻訳テストコレクションによる評価の妥当性を検証する．

本研究では，実験用のテストコレクションとして，多言語質問応答データセットである MKQA [10] を Wikipedia と組み合わせることで構築した英語・日本語間で同一内容とみなせるテストコレクションを用いた．評価実験では疎検索モデルである BM25 といくつかの密検索モデルを対象とし，人手テストコレクションと機械翻訳テストコレクションの評価結果を比較した．比較にあたっては，検索モデルの順位の一致度を

Kendall's $\tau$ により，各クエリにおける文書ランキングの一致度を Rank-Biased Overlap により評価した．さらに，翻訳品質の影響を分析するため，Google 翻訳と 6 種類の公開翻訳モデルを用い，翻訳品質については BLEU により概算した．

本研究の実験結果から，機械翻訳テストコレクションは，検索モデルの順位の類似度という観点では，人手で構築されたテストコレクションと概ね一貫した結論を与えることが確認された．一方で，翻訳品質の違いが評価結果に与える影響については，極端に翻訳品質が低い場合を除き，翻訳品質の高さが妥当性の向上に直結するとは限らないことが示された．

本研究における貢献を以下に示す：

1. 英語・日本語間で対応するテストコレクションを用い，同一言語・同一内容の人手テストコレクションと機械翻訳テストコレクションを比較することで，機械翻訳テストコレクションによる評価の妥当性を検証した．
2. MKQA と Wikipedia を組み合わせることで，情報検索評価に利用可能な英語・日本語間で同一内容とみなせるテストコレクションを構築し，本研究の実験で利用した．
3. Google 翻訳と 6 種類の公開翻訳モデルによる機械翻訳結果を用いて，翻訳品質と機械翻訳テストコレクションによる評価の妥当性との関係を分析した．

本論文の構成は次の通りである． 第 2 節では，本研究と関連する既存研究について述べる．第 3 節では，構築したテストコレクションおよび実験設定について説明する． 第 4 節では，実験結果を示し，機械翻訳テストコレクションの妥当性について考察する．第 5 節では，今後の課題と共に本研究の結論を述べる．

## 2 関 連 研 究

日本語の情報検索評価のためのテストコレクションとしては，これまでにいくつかの取り組みが存在する．例えば，日本語を対象とした情報検索評価の枠組みとして，国立情報学研究所が主催する NTCIR プロジェクトにおいて，これまでに複数のテストコレクションが構築されてきた [7]．一方で，NTCIR におけるテストコレクションは，タスクや対象ドメインが多様であるが，利用にあたっては申請手続きが必要であり，研究目的での利用が前提とされている．また，多言語を対象とした情報検索評価用テストコレクションにおいても，日本語を対象言語の一つとして含むものが提案されている．代表的な例として，MIRACL [22] や Mr.TyDi [21] が挙げられる．近年では，日本語のテキスト埋め込みの評価を目的としたベンチマークとして，JMTEB[1]が公開されている．このベンチマークは，検索を含む複数の評価タスクを対象としており，その一部として，上述の MIRACL や Mr.TyDi を含む既存の検索評価用テストコレクションが収録されている．一方で，JMTEB に含まれるその他のテストコレクションには，クエリの生成方法や適合性判定の方法といった点で多様な背景を持つため，各テストコレクションの特性を考慮することが重要である．このように，日本語においては，英語における BEIR [16] に代表されるような，広く利用される汎用的な情報検索ベンチマークは十分に整備されていない．

こうした状況は日本語に限ったものではなく，英語以外の言語においても高品質なテストコレクションの不足が指摘されている [11], [20]．この課題に対する 1 つのアプローチとして，既存の英語テストコレクションを他言語へ機械翻訳して利用する試みが行われている．Wojtasik らは，英語における大規模な情報検索ベンチマークである BEIR をポーランド語に機械翻訳することで，ポーランド語における情報検索評価のためのベンチマーク BEIR-PL を構築した [20]．同研究では，翻訳後のテストコレクションを用いて多数の検索モデルを対象とした評価実験を行い，モデル性能に関する結果を報告している．一方で，翻訳後のテストコレクションによる評価の妥当性については明示的な検証は行われていない．同様に Lotfi らは，BEIR をオランダ語に機械翻訳することで，オランダ語におけるベンチマーク BEIR-NL を構築した [11]．同研究では，翻訳後のテストコレクションを用いて多数の検索モデルを対象とした評価実験の結果を報告している．また，一部のモデルを用いた BEIR や BEIR-PL による評価結果との比較や，BEIR-NL を英語に逆翻訳したベンチマークと元の英語 BEIR との比較により，機械翻訳という構築方法の評価結果への影響を部分的に検証している．しかし，この分析は，機械翻訳後のテストコレクションが，同一言語における人手のテストコレクションと同様の評価を行えるかを検証したものではない．また，Jeronymo らは，代表的な情報検索評価ワークショップである TREC において構築された英語のアドホック検索用テストコレクションである Robust04 を機械翻訳することで，mRobust04 を構築した [6]．同研究でも，翻訳後のテストコレクションを用いて複数の検索モデルを対象とした評価実験の結果を報告している．ただし，その評価の妥当性についての明示的な検証は行われていない．

テストコレクションが十分に存在しない状況に対する別のアプローチとして，大規模言語モデルを用いて適合性判定を行う手法が提案されている．いわゆる LLM-as-a-Judge は，クエリと検索結果文書を入力として，大規模言語モデルによる適合性判定を行うことで，人手による適合性判定を代替することを目的としている [4], [23]．加えて，大規模言語モデルを用いてクエリおよび適合性判定結果を事前に生成し，人手に依らないテストコレクションを構築する手法として，合成テストコレクションが提案されている．Rahmani らは，構築した合成テストコレクションと人手により構築されたテストコレクションを比較し，評価結果の類似性を分析している [14]．こうした手法は，人手による適合性判定を用いずに評価を行える利点を持つ一方で，生成手法や用いるモデルに依存した特性を持つ可能性があり，評価結果に影響を与え得る体系的なバイアスが生じることも報告されている [15]．

表 1 に示すように，テストコレクション不足への対処法には，機械翻訳テストコレクション，LLM-as-a-Judge，合成テストコレクションといった手法が存在し，それぞれ前提とするデータや適合性判定の生成方法において異なる特性を持つ．

1：https://huggingface.co/datasets/sbintuitions/JMTEB

表 1: テストコレクション不足への対処法の比較

| 手法 | 前提とするデータ | 適合性判定 |
|---|---|---|
| 機械翻訳テストコレクション | **英語** 文書・クエリ・適合性判定 | 人手 |
| LLM-as-a-Judge | **日本語** 文書・クエリ | LLM |
| 合成テストコレクション | **日本語** 文書 | LLM |

LLM-as-a-Judge や合成テストコレクションについては，人手により構築されたテストコレクションと類似した評価結果が得られることが報告されている一方で，生成手法や用いるモデルに依存する可能性も指摘されており，その妥当性については引き続き検討が行われている．

機械翻訳テストコレクションは，既存のテストコレクションに含まれるクエリおよび文書を機械翻訳によって他言語へ変換し，翻訳前のテストコレクションにおける適合性判定との対応付けを保ったまま構築されるテストコレクションである．このような方法は，人手による適合性判定を再利用できるという実用的な利点を持つ一方で，機械翻訳を介した場合に評価結果がどの程度変化するのかについては，十分に明らかになっていない．本研究では，この機械翻訳テストコレクションに着目し，人手により構築された同言語・同一内容のテストコレクションとの比較を通じて，評価結果の関係性を分析する．

## 3 実験設定

本節では，本研究における実験設定について述べる．まず，日英間で同一の内容を持つテストコレクションの構築の概略について説明する．次に，機械翻訳によるテストコレクションの生成の概要を述べる．続いて，本研究で使用した検索モデル，評価方法について述べる．

### 3.1 英語・日本語間で対応するテストコレクションの構築

英語・日本語間で対応する情報検索テストコレクションは，既存には公開されていない．そのため，機械翻訳テストコレクションと人手で構築されたテストコレクションを同一内容の条件下で比較することは困難である．本研究ではこの課題に対処するため，英語と日本語の間で対応付けられたクエリ集合，文書集合，および適合性判定からなる本研究の目的に即したテストコレクションを構築した．

本研究では，異なる言語間で同一の内容を表すクエリ集合を得るため，多言語質問応答データセットである MKQA [10] を用いた．MKQA は，英語の質問応答データセットである Natural Questions [8] を起点として，その英語の質問文を人手により多言語へ翻訳することで構築されたデータセットであり，英語と日本語を含む 26 言語に対応している．本研究では，このうち英語および日本語の質問文を，テストコレクションにおけるクエリとして利用した．

文書集合には Wikipedia を用いた．これは，MKQA の元となっている Natural Questions が Wikipedia の記事を情報源として構築されているためであり，質問応答データを情報検索タスクへ変換する際に一般的に用いられている設計と整合している．本研究では，英語版および日本語版 Wikipedia の記事データを対象とし，2025 年 7 月 20 日版のダンプを利用した．

表 2: 構築したテストコレクションの統計

| 言語 | クエリ数 | 文書数 | 適合性判定数 |
|---|---|---|---|
| 英語 | 1,263 | 663,609 | 1,263 |
| 日本語 | 1,263 | 669,824 | 1,263 |

文書集合の対応付けには，Wikipedia における言語間リンクを用いた．言語間リンクは，異なる言語版における同一概念の記事同士を結び付けるものである．抽出前の文書数は，英語版 Wikipedia で 7,009,646 件，日本語版 Wikipedia で 1,459,624 件であった．これらの中から，英語版の記事のうち日本語記事への言語間リンクを持つもの，および日本語版のうち英語版への言語間リンクを持つものをそれぞれ抽出した．その結果，英語では 663,609 件，日本語では 669,824 件の文書が得られた．Wikipedia の記事は，言語間リンクで結ばれていても，内容が必ずしも一致するとは限らない．そこで本研究では，文書内容の差異による影響を緩和するため，各記事の冒頭のセクションのみを利用した．冒頭のセクションは記事全体の概要を含むことが多く，全文を用いる場合と比較して，内容の差異が小さいと期待され，言語間の内容差による影響を抑制できると考えられる．次に，文書集合への制約に伴い，利用するクエリおよび適合性判定結果を限定した．MKQA には 10,000 件のクエリと適合性判定結果が含まれているが，言語間リンクにより結ばれた文書中に適合文書が存在するクエリのみを抽出した．結果として，5,221 件のクエリが得られた．加えて，文書を冒頭セクションに限定したことにより，その文書を適合文書と判断する根拠となる情報が失われる影響を抑えるため，MKQA における解答が文書中に直接含まれる場合のみを適合文書として採用した．この条件で適合文書をもつクエリは 1,263 件であった．以上の手順より，日英で対応するクエリ集合，文書集合および適合性判定結果からなるテストコレクションを構築した．

表 2 に，構築したテストコレクションの統計を示す．文書集合のサイズは言語間で完全には一致していないが，後述する機械翻訳テストコレクションの生成および検索評価では，適合文書と同一の手続きで抽出したランダム文書を用いて文書集合を構成しており，この不均衡は評価結果に影響しない設計となっている．

### 3.2 機械翻訳テストコレクションの生成

本節では，3.1 節で構築した英語テストコレクションを機械翻訳することで，機械翻訳による日本語テストコレクションを生成する手順について述べる．翻訳対象としたのは，英語テストコレクションに含まれる全てのクエリ，それらに対応する適合文書，および検索評価に用いるためにランダムに抽出した 50,000 件の文書である．適合文書に加えてランダム文書を翻訳対象に含めることで，全文書を翻訳することなく，検索評価に必要な数の非適合文書を含む文書集合を構成した．その結果，機械翻訳テストコレクションは，1,263 件のクエリ，51,263 件

表 3: Tatoeba データセットにおける翻訳品質評価

| 翻訳モデル | BLEU |
|---|---|
| Google 翻訳 | 30.90 |
| M2M-100 (1.2B) | 21.13 |
| NLLB-200 (3.3B) | 20.14 |
| mBART-50 | 19.90 |
| M2M-100 (418M) | 18.85 |
| NLLB-200 (600M) | 17.35 |
| OPUS-MT | 0.88 |

の文書，1,263 件の適合性判定から構成されており，人手テストコレクションと対応関係を持つ．

機械翻訳には，Google 翻訳 [2]，NLLB-200 (3.3B[3]，600M[4]) [3]，M2M-100 (1.2B[5]，418M[6]) [5]，mBART-50[7] [9]，および OPUS-MT[8] [17], [18] を用いた．Google 翻訳は，既存の機械翻訳による情報検索データセットの構築において広く利用されている翻訳サービスであり [2], [6], [20]，本研究においても代表的な翻訳手法として採用した．加えて，翻訳モデルの違いが検索評価に与える影響を確認するため，多言語翻訳を目的として学習された複数の公開モデルを用いた．

### 3.3 翻訳モデルの品質評価

本研究では，機械翻訳テストコレクションの生成に複数の翻訳モデルを用いているため，検索評価実験に先立ち，各翻訳モデルの翻訳品質を自動評価指標により比較した．翻訳品質の評価には，mMARCO [2] における翻訳品質評価の設定に倣い，Tatoeba データセット [9]を用いて BLEU スコア [12] を算出した．具体的には，Tatoeba データセットの 2023 年 4 月 12 日版に含まれる英語と日本語の翻訳ペアの test 分割から，1,000 件をランダムに抽出し，各翻訳モデルによる翻訳結果と参照訳との間で BLEU スコアを計算した．BLEU の算出には SacreBLEU [13] を用い，日本語のトークナイザとして ja-mecab を指定した．

その結果は表 3 に示す通りであり，Google 翻訳が最も高い BLEU スコアを示した．次いで，M2M-100 (1.2B)，NLLB-200 (3.3B)，mBART-50，M2M-100 (418M)，NLLB-200 (600M) の順となった．一方，OPUS-MT は他のモデルと比較して極端に低い BLEU スコアを示した．

### 3.4 検索モデル

本節では，構築・生成した各テストコレクションに対して検索評価を行うために用いた検索モデルについて述べる．いずれの検索モデルも事前学習済みモデルをそのまま使用し，追加の学習やチューニングは行っていない．

2：https://cloud.google.com/translate
3：https://huggingface.co/facebook/nllb-200-3.3B
4：https://huggingface.co/facebook/nllb-200-distilled-600M
5：https://huggingface.co/facebook/m2m100_1.2B
6：https://huggingface.co/facebook/m2m100_418M
7：https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt
8：https://huggingface.co/Helsinki-NLP/opus-mt-en-jap
9：https://tatoeba.org

a） 疎検索モデル

疎検索モデルとして，BM25 を用いた．BM25 は，単語の出現頻度に基づく語彙一致型の情報検索モデルであり，情報検索分野において広く利用されている．実装には Pyserini を用い，Lucene に基づく BM25 を使用した．ハイパーパラメータは Pyserini のデフォルト設定に従い，$k_1 = 0.9$，$b = 0.4$ とした．

b） 密検索モデル

密検索モデルとして，多言語情報検索において近年広く利用されている事前学習済み埋め込みモデルを複数用いた．使用したモデルは，mDPR，mContriever，LaBSE，mE5 (small, base, large)，mGTE，jina-embeddings-v3，bge-m3 である．実装については，mDPR は MIRACL [22] などで用いられている MS MARCO [1] によって事前学習されたモデルを用いた．その他のモデルには Sentence-Transformers を利用した．いずれのモデルについても，事前学習済みモデルをそのまま使用してゼロショットで検索評価を行った．

### 3.5 評価方法

本節では，人手で構築されたテストコレクションと機械翻訳テストコレクションにおける検索評価結果を比較するために用いた評価方法について述べる．

#### 3.5.1 検索性能の評価指標

検索結果の性能を評価する指標として，Recall@100 と nDCG@10 を用いた．Recall@100 は，上位 100 件の検索結果に適合文書が含まれるかどうかを評価する指標であり，適合文書を検索結果中に発見できたかどうかを測るために用いた．nDCG@10 は，検索結果上位において適合文書がどの位置に現れるかを考慮した評価指標である．本研究で用いたテストコレクションでは，各クエリに対する適合文書は高々1 件であり，適合性は 2 値で与えられているため，nDCG@10 は検索結果上位 10 件以内に適合文書が出現するかどうか，およびその順位を反映した指標となる．

#### 3.5.2 評価結果の類似度指標

人手のテストコレクションと機械翻訳テストコレクションに基づく検索評価結果の類似度を分析するため，本研究では順位付けの類似度を測る指標を用いた．評価結果の比較は，検索モデルの順位と，各クエリにおける文書ランキングという 2 つの異なる粒度で行う．

a） 検索モデルの順位の類似度

検索モデルの順位の類似度を評価するため，Kendall's $\tau$ を用いた．Kendall's $\tau$ は，2 つの順位付けの間の相関を測る指標であり，順位の全体的な一致度を評価できる．本研究では，各テストコレクションにおいて得られた検索モデルの評価値に基づき，モデル順位を作成し，人手のテストコレクションと機械翻訳テストコレクションとの間での Kendall's $\tau$ を算出した．なお，Kendall's $\tau$ により得られた順位相関の統計的有意性を検討するため，統計的検定を行った．

b） 文書ランキングの類似度

一方，各クエリにおける文書ランキングの類似度を評価するため，Rank-Biased Overlap (RBO) [19] を用いた．RBO は，

2 つの文書ランキング $S$ および $T$ に対して，ランキングの深さ $d$ を順に増やしながら，上位 $d$ 件に含まれる文書の一致度を用いて類似度を評価する指標である．文書ランキングの順序関係を一様に評価する Kendall's $\tau$ とは異なり，検索結果として重要な上位順位の一致を重視できる点に特徴がある．RBO は次式で定義される：

$$\mathrm{RBO}(S,T,p) = (1-p)\sum_{d=1}^{\infty} p^{d-1} A_d, \tag{1}$$

$$A_d = \frac{|S_{1..d} \cap T_{1..d}|}{d}, \tag{2}$$

ここで $S_{1..d}$ および $T_{1..d}$ は，それぞれ文書ランキング $S$ および $T$ の上位 $d$ 件を表す．$A_d$ は，順位 $d$ までに含まれる文書集合の一致度，すなわち上位 $d$ 件に含まれる共通文書の割合を意味する．$p$ は上位順位をどの程度重視するかを制御するパラメータであり，本研究では RBStar[10]のデフォルトのパラメータである $p = 0.95$ を用いた．

本研究では，各クエリについて上位 1,000 位までの文書ランキングを得ているが，上位の順位が検索結果として特に重要であることから，RBO を用いて文書ランキングの類似度を分析した．

## 4 実験結果

本節では，本研究で実施した検索評価実験の結果を示す．機械翻訳テストコレクションの妥当性を検証するため，本節では 2 つの観点から分析を行う．まず，人手で構築されたテストコレクションと機械翻訳テストコレクションに基づく検索評価結果を比較し，検索モデルの順位の一致度に基づいて妥当性を分析する．次に，翻訳モデルごとの評価結果を比較し，翻訳品質が評価結果に与える影響を分析する．

### 4.1 検索モデルの順位の一致度に基づく妥当性分析

本節では，機械翻訳テストコレクションに基づく検索評価結果が，人手で構築されたテストコレクションによる評価結果と同様の検索モデルの順位を与えるかを分析する．

まず，人手テストコレクションと機械翻訳テストコレクションに基づく検索評価値の対応関係を確認するため，散布図による比較を行った．図 1a および図 1b は，それぞれ Recall@100 および nDCG@10 における，各検索モデルの評価値を，人手テストコレクション（横軸）と機械翻訳テストコレクション（縦軸）で対応付けたものである．これらの散布図から，OPUS-MT を除く多くの翻訳モデルにおいては，機械翻訳テストコレクションに基づく評価値が，人手テストコレクションによる評価値と概ね単調な関係を保っていることが確認できる．一方で，OPUS-MT による機械翻訳テストコレクションでは，検索性能の評価結果が著しく低下しており，他の翻訳モデルとは異なる傾向を示す．

次に，評価値そのものではなく，検索モデル間の相対的な順位関係に着目し，人手テストコレクションと機械翻訳テストコレクションに基づく評価結果の一致度を Kendall's $\tau$ により定量的に評価した．表 4 は，各翻訳モデルを用いた機械翻訳テストコレクションと人手テストコレクションとの間で算出した Kendall's $\tau$ を示している．その結果，OPUS-MT を除く全ての翻訳モデルにおいて，Recall@100，nDCG@10 のどちらの指標においても概ね高い Kendall's $\tau$ の値が得られた．また，これらの順位相関はいずれも有意水準 0.05 において有意であり，人手テストコレクションと機械翻訳テストコレクションの間に統計的に有意な相関が認められた．一方，OPUS-MT については，Kendall's $\tau$ の値が低く，統計的に有意な相関は認められなかった．

表 4: 人手テストコレクションと機械翻訳テストコレクション間のモデル順位の Kendall's $\tau$

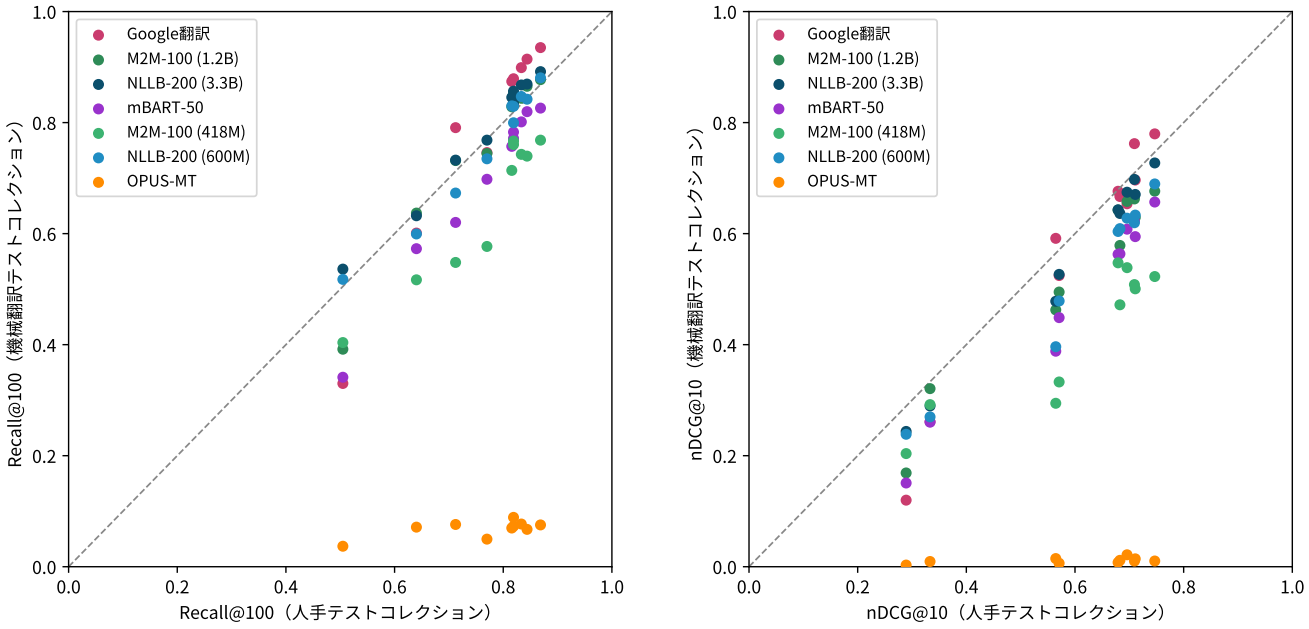
| 翻訳モデル | Recall@100 | nDCG@10 |
|---|---|---|
| Google 翻訳 | 0.8989 | 0.7778 |
| M2M-100 (1.2B) | 0.9439 | 0.8222 |
| NLLB-200 (3.3B) | 0.9439 | 0.8667 |
| mBART-50 | 0.9888 | 0.9111 |
| M2M-100 (418M) | 0.7641 | 0.6000 |
| NLLB-200 (600M) | 0.8540 | 0.9556 |
| OPUS-MT | 0.3146 | 0.3778 |

なお，テストコレクションの代替可能性を検索システム順位の Kendall's $\tau$ により評価した先行研究として，Rahmani らは，クエリおよび適合性判定の双方を LLM により合成したテストコレクションと人手テストコレクションとの間で，nDCG@10 において $\tau = 0.8568$ の一致度を示している [14]．また，Faggioli らは，適合性判定のみを LLM により代替した設定において $\tau = 0.86$ の一致が得られることを示している [4]．これらの研究とは実験設定が異なるものの，OPUS-MT を除く本研究で得られた Kendall's $\tau$ の値は，先行研究において高い一致度が示されている水準と同程度であると言える．以上より，本研究の実験設定においては，一定の翻訳品質を持つ翻訳モデルを用いた場合，機械翻訳テストコレクションは検索モデルの順位の評価という観点から，人手テストコレクションの代替として妥当な評価結果を与えることが示唆された．

### 4.2 翻訳品質の違いが評価結果に与える影響

前節では，検索モデルの順位の一致度に着目し，機械翻訳テストコレクションが人手テストコレクションと高い一致度を示すことを確認した．一方，機械翻訳テストコレクションは翻訳モデルに依存して構築されるため，翻訳モデルの違いが評価結果にどのような影響を与えるかを，より詳細に分析する必要がある．表 3 で示した翻訳品質評価では，Google 翻訳，M2M-100 (1.2B)，NLLB-200 (3.3B)，mBART-50，M2M-100 (418M)，NLLB-200 (600M)，OPUS-MT の順で BLEU スコアが低下することが確認された．本節では，この翻訳品質の違いが検索評価結果にどのような影響を与えるかを分析する．

10：https://github.com/rankbiased/rbstar




(a) Recall@100　　(b) nDCG@10

図 1: 機械翻訳テストコレクションと人手テストコレクションによる検索性能評価結果

表 4 で示した翻訳モデル別の Kendall's $\tau$ に着目すると，OPUS-MT を除くいずれの翻訳モデルにおいても，検索モデルの順位の一致度は高い水準にあることが分かる．しかし，翻訳モデル間で Kendall's $\tau$ の値には差が見られ，その順序関係は翻訳品質評価の結果とは必ずしも一致しない．例えば，BLEU スコアが最も高い Google 翻訳よりも，mBART-50 の方が高い Kendall's $\tau$ を示している．このことから，翻訳品質の自動評価指標による優劣が，検索評価結果の一致度に単純には反映されないことが示唆される．

もっとも，Kendall's $\tau$ は検索モデルの順位関係という比較的粗い粒度での一致度を評価する指標であり，翻訳モデルの違いが検索評価結果に与える影響を十分に捉えきれていない可能性がある．そこで次に，各クエリにおいて検索により得られた文書ランキングの一致度に着目し，Rank-Biased Overlap (RBO) を用いた分析を行う．

表 5 は，翻訳モデル別に，各検索モデルにおけるクエリごとの文書ランキングの RBO の平均を示したものである．全ての検索モデルにおいて，Google 翻訳が最も高く，OPUS-MT が最も低い RBO を示している．一方，その他の中間的な翻訳品質を持つ翻訳モデル間では，RBO の順序が BLEU スコアの順序と必ずしも一致していない．さらに，各検索モデルごとに，翻訳モデルを要因とした Tukey HSD 検定を行った結果，有意水準 0.05 において，OPUS-MT は全ての検索モデルにおいて全翻訳モデルとの間に有意差が認められた．Google 翻訳についても同様に，他の全翻訳モデルとの間に有意差が認められた．一方，その他の中間的な翻訳品質を持つ翻訳モデル同士では，有意差が認められないペアが複数存在した．

BLEU スコアと RBO の関係をより直接的に確認するため，各翻訳モデルの BLEU スコアと全検索モデルにおける RBO 平均の間の Pearson の相関係数 $r$ を算出した．全 7 モデルを対象とした場合には $r = 0.953$ と強い正の相関が認められたが，OPUS-MT を除外した場合には $r = 0.794$ となり，有意水準 0.05 において有意な相関は認められなかった．

これらの結果から，極端に翻訳品質が低い場合には RBO が著しく低下する一方，一定の翻訳品質を持つ翻訳モデル間では，BLEU スコアの高さが RBO の向上に直結するとは限らないことが示された．

#### 4.2.1　低い RBO を示すクエリに関する定性的分析

本節では，RBO が特に低いクエリについて行った定性的な観察の結果を述べる．なお，ここでは翻訳モデル間の比較を目的とするものではなく，BM25 および mE5-large の両検索モデルにおいて，Google 翻訳を用いた場合に RBO が下位に位置したクエリを対象とした．

これらのクエリを確認した結果，多くのクエリにおいて人名や作品名などの固有名詞が含まれており，クエリと文書の間で当該固有名詞の翻訳結果が一致していない例が観察された．例えば，楽曲 Seasons in the Sun に関するクエリでは，クエリ中では「太陽の季節」と翻訳された一方，適合文書中では「シーズンズ・イン・ザ・サン」という表記が用いられていた．このように，検索タスクにおいて重要な手がかりとなる固有名詞について，クエリと文書で使用される語彙が一致していない場合，両者の対応関係が弱まり，文書ランキングの一致度が低下する可能性がある．また，固有名詞を含む英語表現が，翻訳過程において固有名詞として適切に解釈されない場合，一般的な語句として処理され，結果として不自然な文法構造の翻訳文が生成されている例も確認された．このような場合には，クエリが検索に適した表現を十分に保持できず，文書側の表現との対応関係がさらに不安定になると考えられる．これらの事例は，機械

表 5: クエリごとの文書ランキングの一致度（RBO）の平均

| 検索モデル | Google | NLLB-200 (3.3B) | M2M-100 (1.2B) | mBART-50 | M2M-100 (418M) | NLLB-200 (600M) | OPUS-MT |
|---|---|---|---|---|---|---|---|
| BM25 | 0.2053 | 0.1627 | 0.1440 | 0.1340 | 0.1025 | 0.1329 | 0.0078 |
| mContriever | 0.1119 | 0.0990 | 0.0759 | 0.0922 | 0.0784 | 0.0887 | 0.0022 |
| LaBSE | 0.1393 | 0.1315 | 0.1232 | 0.1178 | 0.1030 | 0.1159 | 0.0060 |
| mDPR | 0.2471 | 0.2124 | 0.1843 | 0.1802 | 0.1377 | 0.1852 | 0.0041 |
| mE5-small | 0.2772 | 0.2345 | 0.1887 | 0.2024 | 0.1591 | 0.2114 | 0.0048 |
| mE5-base | 0.2865 | 0.2518 | 0.2126 | 0.2231 | 0.1754 | 0.2327 | 0.0065 |
| mE5-large | 0.3460 | 0.2982 | 0.2519 | 0.2666 | 0.2032 | 0.2703 | 0.0059 |
| mGTE | 0.3366 | 0.2936 | 0.2649 | 0.2648 | 0.2099 | 0.2686 | 0.0093 |
| jina-embeddings-v3 | 0.3544 | 0.2930 | 0.2726 | 0.2518 | 0.2150 | 0.2650 | 0.0078 |
| bge-m3 | 0.3150 | 0.2616 | 0.2299 | 0.2341 | 0.1677 | 0.2268 | 0.0052 |

翻訳による意味的な正確さとは独立に，クエリと文書の間で語彙や表記の一貫性が保持されない場合，文書ランキングの一致度が低下し得ることを示している．したがって，RBO に基づく評価結果を解釈する際には，このような翻訳上の特性を考慮する必要がある．

以上の分析から，一定の翻訳品質を持つ翻訳モデルを用いた場合，機械翻訳テストコレクションは検索モデルの順位という観点では人手テストコレクションと概ね一貫した評価結果を与えることが確認された．一方で，翻訳品質の高さが評価結果の妥当性に直結するとは限らず，Kendall's $\tau$ および RBO のいずれにおいても，翻訳品質の自動評価指標の順序と評価結果の一致度の間に単純な対応関係は認められなかった．

## 5 結　　論

本研究では，日本語検索タスクにおける機械翻訳テストコレクションの妥当性を検証することを目的とし，人手で構築されたテストコレクションと機械翻訳テストコレクションに基づく検索評価結果を比較した．

まず，検索モデルの順位の一致度に基づく分析から，極端に翻訳品質が低い OPUS-MT を除き，機械翻訳テストコレクションは人手テストコレクションとの Kendall's $\tau$ において高い値を示すことが確認された．この結果は，一定の翻訳品質を持つ翻訳モデルを用いた場合，検索モデル間の相対的な性能関係という観点では，人手テストコレクションと概ね整合する評価結果が得られることを示している．

一方で，翻訳品質の高さが評価結果の妥当性に直結するとは限らないことも明らかとなった．Kendall's $\tau$ および RBO のいずれにおいても，翻訳品質の自動評価指標の順序と評価結果の一致度の間に単純な対応関係は認められなかった．

本研究で扱った機械翻訳テストコレクションは日本語のものに限られており，他言語においても同様の傾向が見られるかは明らかではない．また，異なる領域の文書や複数の適合文書を含む設定において，機械翻訳モデルがどのように影響するかも未知数である．本研究で用いたテストコレクションの限界を踏まえ，多言語・多領域での検証が今後の課題として挙げられる．

## 謝　　辞

本研究は JSPS 科研費 JP23K28090 の助成，および情報・システム研究機構 “戦略的研究プロジェクト” の支援を受けたものです．ここに記して謝意を表します．

## 文　　献

[1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2018.

[2] Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. mMARCO: A multilingual version of the MS MARCO passage ranking dataset. *arXiv preprint arXiv:2108.13897*, 2021.

[3] Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and NLLB Team. No language left behind: Scaling human-centered machine translation. *Nature*, Vol. 630, pp. 841–846, 2024.

[4] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 39–50, 2023.

[5] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, Vol. 22, pp. 1–48, 2021.

[6] Vitor Jeronymo, Mauricio Nascimento, Roberto Lotufo, and Rodrigo Nogueira. mRobust04: A multilingual ver-

sion of the TREC Robust 2004 benchmark. *arXiv preprint arXiv:2209.13738*, 2022.

[7] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. Overview of IR tasks at the first NTCIR workshop. In *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 11–44, 1999.

[8] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 452–466, 2019.

[9] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 726–742, 2020.

[10] Shayne Longpre, Yi Lu, and Joachim Daiber. MKQA: A linguistically diverse benchmark for multilingual open-domain question answering. *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 1389–1406, 2021.

[11] Ehsan Lotfi, Nikolay Banar, and Walter Daelemans. BEIR-NL: Zero-shot information retrieval benchmark for the Dutch language. In *Proceedings of the 18th Workshop on Building and Using Comparable Corpora (BUCC)*, pp. 36–45, 2025.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318, 2002.

[13] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, 2018.

[14] Hossein A. Rahmani, Nick Craswell, Emine Yilmaz, Bhaskar Mitra, and Daniel Campos. Synthetic test collections for retrieval evaluation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2647–2651, 2024.

[15] Hossein A. Rahmani, Varsha Ramineni, Emine Yilmaz, Nick Craswell, and Bhaskar Mitra. Towards understanding bias in synthetic data for evaluation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5166–5170, 2025.

[16] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[17] Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez, and Sami Virpioja. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, Vol. 58, pp. 713–755, 2023.

[18] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 479–480, 2020.

[19] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, Vol. 28, No. 4, pp. 1–38, 2010.

[20] Konrad Wojtasik, Kacper Wołowiec, Vadim Shishkin, Arkadiusz Janz, and Maciej Piasecki. BEIR-PL: Zero-shot information retrieval benchmark for the Polish language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2149–2160, 2024.

[21] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. TyDi: A multilingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 127–137, 2021.

[22] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, Vol. 11, pp. 1114–1131, 2023.

[23] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.