インストラクションと複数タスクを利用した日本語向け分散表現モデルの構築

勝又 智　木村 大翼　西鳥羽 二郎
株式会社レトリバ
{satoru.katsumata,daisuke.kimura,jiro.nishitoba}@retrieva.jp

概要

Information Retrieval を始め，様々な用途で分散表現を用いた研究や製品が開発されており，高性能な分散表現モデルへの需要は日々高まっている．本研究では，日本語の多様なタスクに対して効果的な分散表現作成モデルを構築した．作成したモデルの検証結果から，学習データとして様々なタスクのデータが分散表現モデルの学習に効果的なこと，英語に関するデータが日本語向け分散表現に効果があることが確認できた．

1　はじめに

文書を対象とした分散表現モデルは研究用途や実応用などの目的で盛んに使用され，高性能な分散表現モデルの需要は日々高まっている．分散表現を利用した Information Retrieval（IR） [1] を始めとして，Summarization [2] や Classification への活用 [3] などが報告されている．このように，文書の分散表現は自然言語処理において重要な要素の一つである．

分散表現モデル構築の研究は，主に英語を焦点に当てて様々な議論がされている．その中でも特に (1) どのデータを使い，(2) どのようなモデルに，(3) どのような学習を施すのか，が日進月歩で報告されている．学習データに関する研究として，Wang ら [4] は Web 上の半構造データから作成した学習データ 270M 件を利用することで，高精度な分散表現モデルを構築できることを報告している．一方で，Su ら [5] や Asai ら [6] は既存の英語タスクのデータを複数利用することで，様々なタスクについて既存研究と比較して高い性能が得られることを報告している．日本語の分散表現モデルにおいては，Tsukagoshi and Sasano [7] が Wang ら [4] と同様の取り組みを行い，88M 件のデータを利用することで高い性能が達成できることを報告している．本研究では，日本語分散表現モデル構築に向けて，**どのデータを使うのか**という問いについて，Su ら [5] や Asai ら [6] と同様の方向性での検証を実施する．

本研究では，日本語に関する様々なデータを収集し，それらを用いることで分散表現モデルの学習が可能か検証した．また，多言語データでの学習により，言語横断による性能向上が報告されている [8] ことを受け，本研究でも英語に関するデータを利用することにより，より高精度な分散表現モデルが構築できるか検証した．

検証の結果，日本語データで学習したモデルがいくつかのタスクで精度が向上することを確認し，英語データを含めて学習することでさらに性能が向上することを確認した．分析から，学習に使用したタスクデータの種類に応じて，改善する評価指標と改悪する評価指標が異なることを確認した．

2　分散表現構築に向けたデータ作成

本研究では，日本語向け分散表現モデル構築に向けて，日本語と英語の既存タスクに関するデータセットから学習データを構築した．§2.1 では既存研究がどのように学習データを構築しているか論じ，§2.2 にて本研究がどのように作成したかを述べる．

2.1　既存研究のデータ作成手法

分散表現モデルの学習に向けて，クエリ q と関連，非関連文書集合 $d^+ \in D^+$，$d^- \in D^-$，クエリ，文書のインストラクション I_q, I_D を組み合わせたデータ (q, D^+, D^-, I_q, I_D) を用意する．Su ら [5] や Asai ら [6] は既存のタスクから (q, D^+, D^-) を作成し，別途対象となるタスクごとに (I_q, I_D) を作成した．

クエリと対応する文書集合．　学習データ構築に向けて，既存のタスクからクエリとそれに関連した文書，関連しない文書を取得する．既存のタスクとして，IR や Question Answering (QA) であれば，明示

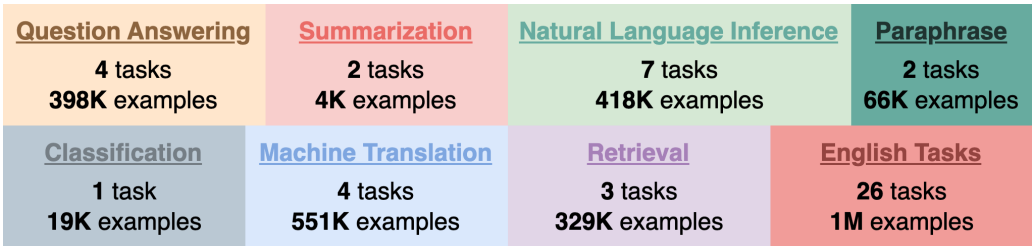

図 1　学習データ作成に使用した日本語データセットの概要.

表 1　本研究と先行研究との概要.

	データセット数	対応言語
Su ら [5]	330 En	En
Asai ら [6]	37 En	En
本研究	19 Ja + 26 En + 4 JaEn	Ja, En

的にこれらの組み合わせが与えられている場合が多いため，その場合は既知の組み合わせを使用した．

一方で，それ以外のタスクではあるクエリに対し，関連文書として何が適切であるかは自明ではない．Su らや Asai らは分類タスクであれば入力テキストをクエリとし，対応するクラスラベルを利用して関連文書を構築する処理を実施した．また，Summarization などの入出力が文または文書の場合には，そのまま入力をクエリとし，出力を関連文書としている．非関連文書については，既存の検索システムを用いて作成している．例として，Asai らは既存のデータセットから作成した文書集合に対して，Contriever [9] を用いて検索を行い，検索結果に対して cross-encoder での類似度フィルタリングを実施して非関連文書を作成している．本研究でもこれらの研究を参考に様々なタスクからクエリと関連，非関連文書の組み合わせ (q, D^+, D^-) を構築した．

インストラクションの作成．　Su らや Asai らはクエリや文書の前に自然言語で記述された**インストラクション**を付与することで，クエリや文書がどのようなものなのかを明示的に表現している．Su らはインストラクションの内容として，(1) 入力がクエリか文書かの記述，(2) タスクの記述，(3) 扱う分野の記述を挙げている．これらを自然文として表現したものをクエリまたは文書の前に付与している．ただし，タスクの記述と扱う分野の記述はデータに応じて付与するかどうかを決めている．本研究でも同様の手順でインストラクションを作成する．

2.2　本研究での学習データ作成手法

本研究では，先行研究と同様の手法を用いて日本語に関するタスクデータから分散表現を作成した．本節では先行研究と異なる点について説明を行う．なお，表 1 に本研究と先行研究の概要を示した．

日本語タスクデータ．　本研究では，日本語タスクデータとして llm-jp-eval [10] や llm-japanese-dataset [11] で利用可能な学習データを選択し，さらに既存の日本語分散表現構築で使用されているデータセット[1)]を利用した．なお，日本語タスクだけでなく，英語タスクも使用することにより，さらに高精度な分散表現モデルが構築可能であるか検証するため，本研究では Asai らが使用した英語データセットの一部を使用している．これらの日英データの言語横断性を向上させるため，本研究では日英 Machine Translation（MT）に関するデータセットも使用している．本タスクで利用した日本語データセットの概要は図 1 に記載し，詳細は §A.1 に記載する．

合成データ．　前述のデータソースでは IR と Classification について不足していたため，本研究では Large Language Model（LLM）を用いて合成データを作成した．IR については，Dai ら [12] の手法を参考に，LLM-jp Corpus v2[2)] の一部の Common Crawl 文書に対して，関連する疑似クエリを生成した．Classification については，2 値の極性レビュー分類データを作成した．具体的には，拡張固有表現分類[3)]を元に，架空の固有表現を作成し，{positive, negative} のどちらかを反映したレビュー文書を作成している．固有表現分類を元にすることで，多様なレビュー文書を生成するようにしている．

非関連文書．　既存のタスクからクエリ q と関連，非関連文書 D^+, D^- の構築は Asai らとおおよそ同じ手法を使用している．違いとして，非関連文書の作成方法が挙げられる．本研究では日本語タスク，英語タスクについて，非関連文書作成の時間効率を考え，Contriver ではなく，bm25 [13] を利用した．その検索結果に対して cross-encoder でのフィル

1)　https://huggingface.co/datasets/hpprc/emb
2)　https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v2
3)　http://ene-project.info/ene9/

表 2　JMTEB での評価結果.

	Classification	Reranking	Retrieval	STS	Clustering	PairClassification	Average
RetrievaBert w/o training	**73.64**	89.88	24.66	67.07	**48.27**	**62.35**	60.98
RetrievaBert w/ 日本語データ	71.99	91.91	59.75	78.32	46.52	62.20	68.62
RetrievaBert w/ 日英混合	71.87	**92.94**	**62.92**	**78.95**	45.85	62.20	**69.12**
`cl-nagoya/ruri-base`	75.49	92.91	69.53	82.87	52.40	62.38	72.60
`intfloat/multilingual-e5-base`	69.86	92.90	69.45	80.45	51.62	62.35	71.11

タリングを実施している．なお，日英 MT データについては，言語横断性のため，bm25 スコアではなく，既存の多言語ニューラル分散表現モデルを利用して候補文書集合を構築している．

インストラクションの作成．　インストラクションについては Su らの手法を参考に，各タスクごとに人手で日本語，英語インストラクションを作成した．ただし，Su らと違い，総データセット数が少なくインストラクションの種類に限りがあるため，本研究では 1 つのインストラクションに対して最大 3 つの言い換えたものを LLM を用いて作成した．学習の際には，最大 4 つのインストラクションからサンプリングされたものが各事例に付与される．

3　分散表現モデルの性能調査

本研究では作成したデータを用いて分散表現モデルが学習可能か，またどの程度の性能なのかを検証した．さらに，日本語タスクデータのみで学習した場合と，日本語英語タスクデータを混合して学習した場合について比較を行い，英語のタスクデータが日本語向け分散表現構築に効果的か検証した．

3.1　実験設定

本研究では単一の Encoder を学習し，そのモデルを用いてクエリと文書のベクトル化を行う．Encoder の学習には in-batch negative を考慮した infoNCE loss [14] を用いた．この学習の詳細については §A.2 に記載する．

学習データ作成の際，特定のデータセットが過剰に偏ることを防ぐため，クエリ数が 1M を超えるデータセットについては 100K まで down-sampling を実施している．作成したデータでの分散表現モデル学習の際に使用したハイパーパラメータや，学習データ作成時に使用した各種学習済みモデルについては §A.3，A.4 に記載する．本研究では分散表現構築の際のベースモデルとして，独自に構築した RetrievaBert[4] の base サイズを使用した．このベースモデルは LLM-jp Corpus v2 と RefinedWeb [15]，The Stack [16]，中国語，韓国語 Wikipedia を利用して学習したものである．

検証タスクとして，本研究では JMTEB[5] を使用した．この JMTEB は Classification，Clustering，IR など合計 6 種類 16 タスクから構築された日本語分散表現ベンチマークである．

比較として，今回の検証モデルと同等のパラメータサイズでの既存モデルも評価を行った．具体的には，Tsukagoshi and Sasano [7] による日本語大規模データを利用した base サイズモデル `cl-nagoya/ruri-base` と，Wang ら [4] の手法を多言語データで実施した base サイズモデル `intfloat/multilingual-e5-base` [17] を検証した．

3.2　JMTEB 実験結果

検証結果を表 2 に示す．本検証の結果から，複数タスクを利用した学習を行うことで，特に Retrieval タスクと STS で性能が向上していることがわかる．一方でいくつかのタスクでは学習することにより性能が減少した．原因調査のため，§4 にて学習データのタスクと JMTEB の評価タスクの関係を調査する．他の学習データ作成手法で訓練されたモデルと比較すると，STS，Clustering について差があることがわかる．この 2 種のタスクについては，細かい粒度での類似度が要求されるタスクと考えられる．提案手法を用いることで Retrieval と STS に対して向上するが，大規模データを利用した場合と比較すると細かい粒度の分散表現構築については未だ改善の余地が存在していると考えられる．

また，学習データとして日本語のみと日英混合を比較した場合，日英混合データを用いることで多くのタスクで性能が向上することがわかる．この結果から，本手法は日本語データだけでなく，英語データを用いることでより高精度な分散表現モデルを構築することができると考えられる．

4）`https://huggingface.co/retrieva-jp/bert-1.3b`

5）`https://github.com/sbintuitions/JMTEB/tree/v1.3.1`
JMTEB の概要は §B に記載.

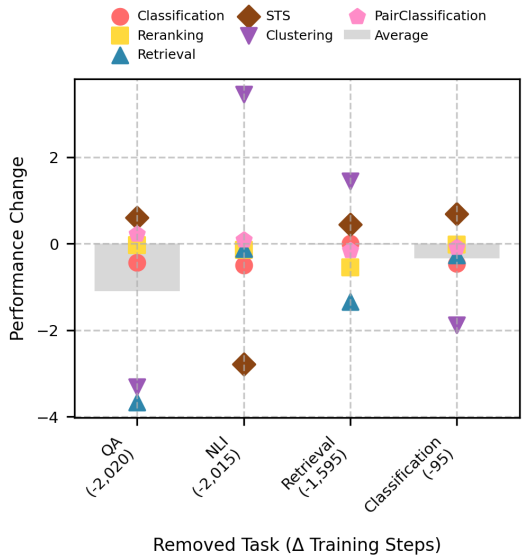

図 2　特定のタスクデータを取り除いた影響.

4　議論

本研究では学習データとして様々なタスクを利用し，それぞれのデータに対してインストラクションを付与した．本節ではこれらの工夫の効果について分析を行う．

Task Ablation Study．　本研究では様々なタスクデータを学習データとして用いた．どのようなタスクが JMTEB に対して効果的か調査を行うため，学習データとして特定のタスクを取り除いた際の影響を調査した．本節では，特定のタスクデータを取り除いたことで性能が減少する事例を **正の影響** と呼称し，その逆を **負の影響** と呼称する．

主要な調査結果を図 2 に記載する．この結果から，QA データは JMTEB の各タスクに対して平均的に正の影響が大きいことがわかる．特に Retrieval タスクと Clustering タスクに対して強い正の影響が確認できる．一方で，Retrieval データは JMTEB の Retrieval タスクに対して正の影響が確認できるが，Clustering タスクについては負の影響となった．また，NLI データは STS タスクに対して強い正の影響が確認できるが，Clustering タスクに対しては負の影響が存在している．これらの結果から，各タスクデータは様々な JMTEB タスクに対して正負の影響を与えていることがわかる．汎用的で高精度な分散表現モデル構築に向けて，これらの影響を適切に考慮したデータ選択の重要性が示唆される．

また，興味深い事柄として，Classification データが JMTEB に対して平均的に正の影響を与えていることがわかる．本研究では Classification データは合成データのみであることを勘案すると，Classification データの重要性が示唆される結果となった．

表 3　インストラクションの検証結果.

学習時	推論時	JMTEB AVG.	FollowIR Benchmark
		68.64	-
✓		68.18	-
	✓	66.92	-4.42
✓	✓	**68.62**	**-1.75**

Instruction Ablation Study．　本研究では学習データにインストラクションを付与した．このインストラクションの有無が性能に対してどのような影響を与えるか調査を行う．評価として，§3.2 で使用した JMTEB と，Weller ら [18] による評価データ（FollowIR Benchmark）を用いた[6]．FollowIR Benchmark はインストラクションが有効か調査するために設計された評価ベンチマークであり，評価尺度として p–MRR と呼ばれる，インストラクションを考慮した検索が行われているか評価を行う尺度を用いる．

調査結果を表 3 に記載する．この結果から，JMTEB については学習時と推論時にインストラクションの有無が一貫していることが重要であることがわかる．一方で，FollowIR Benchmark を確認すると，インストラクションを含めて学習したモデルは推論時もインストラクションを考慮した検索を行っていると考えられる．以上の結果から，インストラクションを考慮する必要があるような柔軟な検索に向けて，学習データに様々なインストラクションを含めることは重要であると考えられる．

5　おわりに

本研究では高精度な分散表現モデルの構築に向け，**どのデータを使うのか**という問いについて検証を実施した．検証の結果，(1) 日本語だけでなく，英語を含めた様々なデータが有効であること，(2) 学習に使用したタスクデータがある評価タスクには有効だが，別の評価タスクでは負の影響を与えること，(3) インストラクションを利用することで，より柔軟な検索が行えることを確認した．課題としていくつかの評価タスクで既存のモデルと差異があることを確認できた．この対策として，学習に使用するタスクをより適切に選択することが考えられる．

6）本データは英語で公開されているため，§A.4 に記載した LLM を用いて日本語に翻訳して使用した．

謝辞

本研究は九州大学情報基盤研究開発センター研究用計算機システムの民間利用を利用したものです.

参考文献

[1] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics.

[2] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6197–6208, Online, July 2020. Association for Computational Linguistics.

[3] Tim Schopf, Daniel Braun, and Florian Matthes. Evaluating unsupervised text classification: Zero-shot and similarity-based approaches. **Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval**, 2022.

[4] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2024.

[5] Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 1102–1121, Toronto, Canada, July 2023. Association for Computational Linguistics.

[6] Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. Task-aware retrieval with instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 3650–3675, Toronto, Canada, July 2023. Association for Computational Linguistics.

[7] Hayato Tsukagoshi and Ryohei Sasano. Ruri: Japanese general text embeddings, 2024.

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[9] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2022.

[10] Namgi Han, 植田暢大, 大嶽匡俊, 勝又智, 鎌田啓輔, 清丸寛一, 児玉貴志, 菅原朔, Bowen Chen, 松田寛, 宮尾祐介, 村脇有吾, 劉弘毅. llm-jp-eval: 日本語大規模言語モデルの自動評価ツール. 言語処理学会第 30 回年次大会, pp. 2085–2089, 2024.

[11] Masanori HIRANO, Masahiro SUZUKI, and Hiroki SAKAJI. llm-japanese-dataset v0: Construction of Japanese Chat Dataset for Large Language Models and its Methodology, 2023.

[12] Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples. In **The Eleventh International Conference on Learning Representations**, 2023.

[13] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In **Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '94, p. 232–241, Berlin, Heidelberg, 1994. Springer-Verlag.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In **Proceedings of the 37th International Conference on Machine Learning**, ICML'20. JMLR.org, 2020.

[15] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon LLM: Outperforming curated corpora with web data only. In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[16] Denis Kocetkov, Raymond Li, Loubna Ben allal, Jia LI, Chenghao Mou, Yacine Jernite, Margaret Mitchell, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro Von Werra, and Harm de Vries. The stack: 3 TB of permissively licensed source code. **Transactions on Machine Learning Research**, 2023.

[17] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report, 2024.

[18] Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. FollowIR: Evaluating and teaching information retrieval models to follow instructions, 2024.

表 4　使用したデータ一覧.

日本語データ	
QA	AutoWikiQA，JEMHopQA，JSQuAD，MQA-ja
NLI	AutoWikiNLI，JaNLI，JNLI，JSICK，JSNLI，Multilingual NLI，NU-MNLI
Paraphrase	PAWS-X，SNOW
Retrieval	MIRACL，JapaneseWikipediaHumanRetrieval，合成データ
Summarization	ParaNatCom，Wikinews
MT	ALT Translation，NLLB，ParaNatCom，Wikipedia's Kyoto Article
Classification	合成データ
英語データ	
QA	GooAQ，HotpotQA，Natural Question，PAQ，PubMedQA，SQuADv2，TriviaQA
NLI	QuoraQuestionPair，SNLI，SPECTER
Paraphrase	Altlex，Medical Text Simplification，MedMCQA，Simple wiki，WikiAnswers
Summarization	CNN/DM，Gigaword，MultiLexSum，scitldr
Others	Qrecc，Wizard of Wiki，fever，Coco captions，Flicker，SentenceCompression，AIDA CoNLL-YAGO

A　詳細な実験設定

本節では使用したデータ，事前学習モデル一覧を示す.

A.1　使用したデータ一覧

本研究では表 4 のデータを用いて分散表現モデルの学習を実施した.

A.2　分散表現の学習

本研究では in-batch negative を考慮した infoNCE loss [14] を最小化するように分散表現を学習した.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\mathrm{sim}(\mathbf{q}_i,\mathbf{d}_i^+)/\tau}}{\sum_{y\in B(i)} e^{\mathrm{sim}(\mathbf{q}_i,y)/\tau}} \tag{1}$$

ただし，N はバッチサイズを表しており，$B(i)$ は以下のように定義される：

$$B(i) = \{\mathbf{d}_i^+\} \cup \{\mathbf{d}^-\} \tag{2}$$

ここで，τ は温度パラメータである.

A.3　分散表現モデル学習時のハイパーパラメータ

本研究では 表 5 記載のハイパーパラメータを用いて学習を実施した.

表 5　ハイパーパラメータ.

GPU	NVIDIA H100 1 枚
バッチサイズ	1024
温度パラメータ	0.05

A.4　使用した事前学習モデル一覧

本研究では表 6 の事前学習モデルを用いた.

表 6　使用した事前学習モデル一覧.

合成データ作成	google/gemma-2-9b-it
非関連文書作成	
候補作成	bm25
cross-encoder	BAAI/bge-reranker-v2-m3
候補作成（MT）	intfloat/multilingual-e5-large
インストラクション作成	google/gemma-2-9b-it
FollowIR 日本語翻訳	cyberagent/calm3-22b-chat

B　JMTEB の概要

JMTEB は日本語向けの分散表現モデル評価ベンチマークである．Classification，Reranking，Retrieval，STS，Clustering，PairClassification の 6 種類から構成されていて，これらは表 7 に記載の 16 タスクから構成されている．評価尺度は各タスクごとに定義されており，すべて大きい方が良いことを示すものとなっている.

表 7　JMTEB 評価タスク一覧.

Classification	AmazonReviewClassification AmazonCounterFactualClassification MassiveIntentClassification MassiveScenarioClassification
Reranking	esci
Retrieval	JAQKET MR.TyDi-ja JaGovFaqs-22k NLP Journal
STS	JSTS JSICK
Clustering	Livedoor-News MewsC-16-ja
PairClassification	PAWS-X-ja