JMedWiC：日本語医療分野の語義同一性判定データセット

堀口 航輝 [1]　杉山 誠治 [1]　梶原 智之 [1,2]　若宮 翔子 [3]　荒牧 英治 [3]
[1] 愛媛大学大学院理工学研究科　[2] 大阪大学 D3 センター　[3] 奈良先端科学技術大学院大学
{horiguchi@ai., sugiyama@ai., kajiwara@}cs.ehime-u.ac.jp
{wakamiya, aramaki}@is.naist.jp

概要

本研究では，医療分野に特化した日本語の語義同一性判定のためのデータセット JMedWiC を構築し，公開する．単語の意味が文脈によって変化する語義曖昧性の問題に対して，先行研究では，所与の 2 文脈における単語の意味が同一か否かを判定する WiC データセットを整備し，語義同一性判定の評価に使用してきた．医療分野においては，語義の誤解釈が医療情報の理解を妨げる可能性があるが，日本語の医療分野に特化した WiC は存在しない．そこで本研究では，大規模コーパスから自動抽出した文脈ペア中の単語に人手で同義性ラベルを付与し，医療分野に特化した日本語の WiC を構築する．

1　はじめに

単語が文脈によって異なる意味を取る語義曖昧性は，機械翻訳 [1] や情報検索 [2] などの下流タスクにおける性能低下の一因である．この問題に対処するために，文脈中の単語に対して適切な語義ラベルを割り当てる Word Sense Disambiguation [3] を中心に，語義曖昧性解消の研究が進められてきた．近年では，語義ラベル体系に依存しない枠組みとして，2 つの文脈における同一単語の意味が同じか否かを判定する Word-in-Context (WiC) [4] が提案され，英語を中心に WiC データセットが整備されている [4–7].

医療分野では，日常的な単語と同一の表記を持ちながら，専門的な文脈では異なる意味を持つ単語が数多く存在する．例えば，「ポケット」という語は，日常会話では衣服などに縫い付けられた袋状の部分を指す一方で，外科領域では皮膚欠損部よりも広い創腔を，歯学領域では歯と歯肉の間に形成される溝を指す．このような医療分野特有の多義性は，既存の WiC データセットでは網羅できていない．そのため，英語では医療分野に特化した WiC データセット [8] が構築されているが，日本語には存在しない．

また，既存の WiC データセットの多くは，WordNet [9] や UMLS [10] などの語義体系を明示的に定義した語彙資源をもとに，対象語の語義が一致する文脈ペアと一致しない文脈ペアを抽出することで構築されている．一方で，語義体系が十分に整備されていない日本語の医療分野においては，同様の手法でデータセットを構築するのは困難である．

本研究では，医療分野における日本語の WiC データセットを構築するために，BERT [11] が生成する文脈化単語分散表現を利用する．文脈化単語分散表現は，同一単語であっても文脈の違いに応じた意味の差異を連続的な表現として捉えることができる．この性質を利用し，擬似的な同義ペアと非同義ペアを自動抽出した上で人手アノテーションを実施し，日本語の医療分野のための JMedWiC[1)] を構築する．なお，JMedWiC は医療分野と一般分野の 2 つのサブセットで構成され，1,000 件ずつの文脈ペアを含む．

2　関連研究

WiC [4] とは，2 つの文脈における同一単語の意味が同じか否かを判定するタスクであり，英語を中心にデータセットが整備されている [4–7]．WiC データセットは，単語を語義単位 (synset) で体系化した語彙資源 WordNet [9] を中核としており，対象語が同一の語義 ID を持つか否かによって構築されている．日本語では，フレーム意味論に基づく語彙資源である日本語フレームネット [12] を基盤とし，対象語が同一フレームに属するか否かに基づいて JWiC [7] が構築されている．医療分野では，医学・生物医学の概念体系を統合した Unified Medical Language System (UMLS) [10] を基盤とし，対象語の概念 ID の一致に基づいて BioWiC [8] が構築されている．

これらの既存データセットはいずれも語義体系を明示的に保持する語彙資源に依存しており，そのような語彙資源を欠く領域では，同様の手法による

1) https://github.com/EhimeNLP/JMedWiC

表 1　JMedWiC の事例

	対象語	文脈 A	文脈 B	ラベル
medical	熱	極端な**熱**や低温から皮膚を守らなければなりません。	病原体は**熱**に弱く界面活性剤により失活する。	True
	胸壁	**胸壁**の良性腫瘍として最も頻度の高いのはどれか。	銀地に、赤い 2 本の**胸壁**のある塔のある城。	False
general	運賃	**運賃**は地下鉄を利用する時間帯による。	**運賃**は往復 3 ドルであった。	True
	足取り	PDF の初期の普及の**足取り**は緩やかなものであった。	得意手は右四つ、押し、**足取り**。	False

データセット構築は困難である．日本語の医療分野も，WordNet や UMLS のような語彙資源は整備されておらず，WiC データセットは構築されていない．

3　JMedWiC の構築

日本語における語義同一性の判定能力を評価するために，WiC 形式のデータセット JMedWiC を構築する．JMedWiC は，医療用語を対象とする medical と，一般用語を対象とする general の 2 つのサブセットで構成される．本研究では，各サブセットにおいて評価対象となる語彙を選定し，それらの出現文脈を 2 種類のテキストコーパスから収集する．そして，対象語の意味的類似度に基づいて文脈ペアを自動抽出し，人手アノテーションを通じて最終的なラベルを決定する．表 1 に JMedWiC の事例を示す．

3.1　対象語彙の選定と文脈抽出

medical では，医療分野の語彙リストとして JMED-DICT[2] [13] を用いる．JMED-DICT は，日本語医療分野の語彙を体系的に収集した辞書であり，BODY（部位），MEDICATION（医薬品），DISEASE（病名）の 3 つのサブデータから構成される．この辞書から，頻度情報に基づいて 9,257 語を抽出し，5 文字以下の 6,476 語を対象語とした．一方 general では，BCCWJ 語彙表[3] の頻度上位から内容語 7,500 語を抽出し，5 文字以下の 7,395 語を対象語とした．

本研究では，幅広い一般的な文脈を含むコーパスとして Wikipedia[4] を利用し，加えて医療分野の文脈を確保するためにオンライン医学事典の MSD マニュアル[5] から医療テキストを収集した．各コーパスに対し，ルールベースの文分割[6] を適用し，medical では Wikipedia および MSD マニュアル，general では Wikipedia から対象語を含む文を抽出した．この際，対象語の出現位置を正確に特定するため，医療辞書データ[7] を組み込んだ MeCab[8] [14] による単語分割を適用した．また，極端に短い文や冗長な文を除外するため，文長が 10 文字以上 50 文字以下の文のみを対象とした．

3.2　文脈ペアリング

人手アノテーションの候補である擬似的な同義ペアと非同義ペアを効率的に抽出するために，文脈化された単語分散表現の余弦類似度に基づく文脈ペアリング手法を提案する．文脈化された単語分散表現は，同一単語であっても文脈によって異なるベクトル表現を持つため，2 つの文脈におけるベクトル間の余弦類似度に基づいて語義の近さを推定できると考えられる．

本手法では，各文脈をマスク言語モデルに入力し，対象語に対応するトークンの隠れ状態ベクトルを文脈化された単語分散表現として用いる．対象語が複数のサブワードに分割される場合は，それらのベクトルの平均を用いる．同一の対象語を含む文脈ペア間でこのベクトルの余弦類似度を計算し，その値が所定の閾値 θ_{high} 以上のペアを擬似的な同義ペア，θ_{low} 以下のペアを擬似的な非同義ペアとして抽出する．

3.3　データセットの自動構築

文脈ペアリングに適したマスク言語モデルを選定するため，単語の意味的な対応関係を同定する能力を測る単語アライメントタスクを用い，東北大 BERT[9] [11]，早大 RoBERTa[10] [15]，ModernBERT[11] [16]，JMedRoBERTa[12] [17] の 4 種類の日本語マスク言語モデルを評価した．評価用デー

2）本研究では無償版である JMED-DICT mini を利用した．https://sip3-d2.naist.jp/jmed-dict.html
3）『現代日本語書き言葉均衡コーパス』短単位語彙表 ver.1.0
4）https://huggingface.co/datasets/range3/wiki40b-ja
5）https://www.msdmanuals.com
6）https://github.com/wwwcojp/ja_sentence_segmenter
7）https://sociocom.naist.jp/j-meddic-for-mecab/
8）https://taku910.github.io/mecab/
9）https://huggingface.co/tohoku-nlp/bert-base-japanese-v3
10）https://huggingface.co/nlp-waseda/roberta-base-japanese
11）https://huggingface.co/sbintuitions/modernbert-ja-130m
12）https://huggingface.co/alabnii/jmedroberta-base-sentencepiece

表 2 SimAlign による単語アライメントの結果

	適合率	再現率	F 値
東北大 BERT	65.51	76.64	**70.64**
早大 RoBERTa	65.63	72.08	68.70
ModernBERT	64.31	76.71	69.96
JMedRoBERTa	53.30	77.76	63.24

タには，意味的に対応する難解文と平易文からなる医療テキスト平易化パラレルコーパス[13] [18] の日本語評価セットから抽出した 300 文対を用いた．各文に対し，医療辞書データ[7]を組み込んだ MeCab [14] による単語分割を適用し，第一著者が人手で単語アライメントを付与した．アライメント手法には 文脈化された単語分散表現の余弦類似度が最大となる単語対を対応付ける SimAlign[14] [19] を用い，F 値により性能を評価した．

表 2 より，東北大 BERT が最も高い F 値を示したため，本研究では東北大 BERT を採用し，3.1 節で抽出した文脈集合に対して 3.2 節の文脈ペアリングを実施した．ここで，medical において非医療分野の文脈同士がペアとなるのを防ぐために，文脈ペアの少なくとも一方が MSD マニュアル由来であることを条件とした．また，文脈ペアリングにおける余弦類似度の閾値は $\theta_{\text{high}} = 0.9$，$\theta_{\text{low}} = 0.6$ と設定した．

文脈ペアリングによって得られた候補集合から，medical では擬似的な同義ペア 700 件，非同義ペア 300 件の計 1,000 件を，general では擬似的な同義ペア 500 件，非同義ペア 500 件の計 1,000 件を選択した．medical において非同義ペアの比率が低いのは，MSD マニュアルのデータサイズが Wikipedia と比較して限られており，抽出元となる文脈数を十分に確保できなかったことに起因する．なお，MSD マニュアルから収集した文脈は再配布の制約があるため，選択した 1,000 件のうち該当する文脈について，医療分野に特化した大規模言語モデル (LLM) である MedQwen-72b[15] [20] を用いて，意味を保持したまま別の表現に書き換えた．

3.4 同義性アノテーション

収集した文脈ペアに対して，medical では医療従事者 1 名，general では学生 1 名が，対象語の語義が同一か否かをアノテーションした．アノテーションの結果，各サブセット 1,000 件からなる WiC 形式のデータセットが構築された（表 3）．余弦類似度の閾値に基づく自動ラベルと人手ラベルの一致率は，medical では 95.8%，general では 85.5%であった．

表 3 JMedWiC のラベル分布（括弧内は自動抽出時の件数を表す．）

	True	False	合計
medical	716 (700)	284 (300)	1,000
general	629 (500)	371 (500)	1,000

4 評価実験

本研究で構築した JMedWiC データセットを用いて，3 種類の教師なし手法の性能を評価する．

4.1 実験設定

マスク言語モデル 各文脈を BERT に入力し，対象語に対応するサブワード埋め込みを平均して文脈化単語分散表現を取得する．2 つの文脈から得られた単語分散表現間の余弦類似度を計算し，類似度が閾値 θ 以上であれば語義が同一，閾値 θ 未満であれば語義が異なると判定する．モデルには，東北大 BERT[9]，早大 RoBERTa[10]，ModernBERT[11]，JMedRoBERTa[12] の 4 種類の日本語 BERT を使用した．閾値は $\theta \in \{0.50, 0.55, \ldots, 0.95\}$ の中から，F 値を最大化する値を採用した．なお，東北大 BERT はデータセット構築時（3.3 節）に使用しており，他のモデルより有利な条件にあることに注意されたい．

マスク言語モデル+クラスタリング 文脈化単語分散表現をクラスタリングする山内ら [21] の手法に倣い，クラスタ数の事前指定が不要な密度ベースのクラスタリング手法である DBSCAN [22] を用いて語義の同一性を評価する．medical では Wikipedia[4] および MSD マニュアル[5]から，general では Wikipedia から対象語が出現するすべての文脈を収集し，文脈化単語分散表現を取得する．そして，評価対象となる文脈を含めた文脈表現集合に対して DBSCAN を適用し，2 つの文脈が同一クラスタに属する場合は語義が同一，異なるクラスタに属する場合は語義が異なると判定する．モデルには，前項と同じく 4 種類の日本語マスク言語モデルを使用した．

大規模言語モデル 大規模言語モデル (LLM) に指示文，対象語，文脈ペアを与える 0-shot の文脈

13） https://github.com/EhimeNLP/MultiMSDcorpus
14） SimAlign のパラメータは token_type=bpe, method=argmax, distortions=0.1 とした．
15） https://huggingface.co/pfnet/Preferred-MedLLM-Qwen-72B

表 4　実験結果

		medical			general		
		適合率	再現率	F 値	適合率	再現率	F 値
マスク言語モデル	東北大 BERT	0.926	0.961	0.943	0.894	0.822	0.857
	早大 RoBERTa	0.929	0.872	0.899	0.827	0.790	0.808
	ModernBERT	0.873	0.887	0.880	0.699	0.882	0.780
	JMedRoBERTa	0.720	0.982	0.831	0.629	0.995	0.771
マスク言語モデル +クラスタリング	東北大 BERT	0.814	0.940	0.872	0.694	0.596	0.642
	早大 RoBERTa	0.722	0.953	0.822	0.714	0.905	0.798
	ModernBERT	0.714	0.982	0.827	0.716	0.996	0.833
	JMedRoBERTa	0.682	0.751	0.715	0.717	0.979	0.828
大規模言語モデル	Swallow-8b	0.929	0.810	0.866	0.950	0.482	0.639
	llmjp-13b	0.954	0.749	0.839	0.972	0.331	0.493
	Qwen-72b	0.940	0.926	0.933	0.960	0.655	0.779
	MedQwen-72b	0.946	0.824	0.881	0.946	0.669	0.784

内学習により，2 つの文脈における対象語の語義が「同じ」か「違う」のいずれかを出力させる．モデルには，汎用 LLM として Swallow-8b[16] [23]，llmjp-13b[17] [24]，Qwen-72b[18] [25]，医療特化 LLM として MedQwen-72b[15] [20] を使用した．

指示文

以下の 2 つの文において対象単語の意味が同じかどうかを判断してください。同じ意味であれば「同じ」、異なる意味であれば「違う」と答えてください。

4.2　実験結果

表 4 に実験結果を示す．東北大 BERT は，データセット構築時と同一の文脈表現に基づいて余弦類似度が計算されるため，いずれのドメインにおいても顕著に高い性能を示した．東北大 BERT を除いて比較すると，medical における評価では汎用 LLM の Qwen-72b，general における評価では DBSCAN を用いた ModernBERT が最高性能を達成した．全体として，余弦類似度の閾値が最適な状態のマスク言語モデルと比較しても，マスク言語モデル+クラスタリングおよび大規模言語モデルが同等またはそれ以上の性能を示した．

ドメイン間で比較すると，medical の性能が general を上回る傾向が確認された．医療用語は専門的な文脈で用いられることが多く，語義の使用範囲が限定されるため，語義判定が容易であった可能性がある．一方，一般用語は多義的あるいは比喩的な用法を含む文脈が多く，語義境界が曖昧になりやすいため，語義判定の難易度が高いと考えられる．

medical において汎用モデルと医療特化モデルを比較すると，医療特化モデルに優位性は見られず，JMedRoBERTa も MedQwen-72b のいずれも汎用モデルの性能を下回った．この結果は，WiC タスクに重要なのは対象語に関する専門知識そのものではなく，文脈から語義の同一性を識別するための汎用的な言語知識である可能性を示唆している．

5　おわりに

本研究では，日本語の医療分野における文脈依存の語義同一性の判定能力を評価するために，文脈化単語分散表現に基づく文脈ペアリングおよび人手アノテーションによって，医療分野と一般分野でそれぞれ 1,000 件からなる JMedWiC を構築した．評価実験の結果，医療分野では大規模言語モデルが，一般分野ではマスク言語モデル+クラスタリングが高い性能を示した．また，医療特化モデルが必ずしも汎用モデルを上回らないことや，ドメイン間でタスクの難易度に違いがあることが明らかになった．

16）https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5
17）https://huggingface.co/llm-jp/llm-jp-3.1-13b-instruct4
18）https://huggingface.co/Qwen/Qwen2.5-72B-Instruct

謝辞

本研究は，戦略的イノベーション創造プログラム（SIP）「統合型ヘルスケアシステムの構築」JPJ012425の助成を受けて実施した．

参考文献

[1] Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word Sense Disambiguation Improves Statistical Machine Translation. In **Proc. of ACL**, pp. 33–40, 2007.

[2] Christopher Stokoe, Michael P. Oakes, and John Tait. Word sense disambiguation in information retrieval revisited. In **Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval**, p. 159–166, 2003.

[3] Roberto Navigli. Word sense disambiguation: A survey. **ACM Computing Surveys**, Vol. 41, No. 2, pp. 10:1–10:69, 2009.

[4] Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In **Proc. of NAACL**, pp. 1267–1273, 2019.

[5] Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization. In **Proc. of EMNLP**, pp. 7193–7206, 2020.

[6] Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. WiC-TSV: An Evaluation Benchmark for Target Sense Verification of Words in Context. In **Proc. of EACL**, pp. 1635–1645, 2021.

[7] 吉田あいり, 河原大輔. 日本語 WiC データセットの構築と読みづらさ検出への応用. 言語処理学会第 29 回年次大会, pp. 1643–1647, 2023.

[8] Hossein Rouhizadeh, Irina Nikishina, Anthony Yazdani, Alban Bornet, Boya Zhang, Julien Ehrsam, Christophe Gaudet-Blavignac, Nona Naderi, and Douglas Teodoro. A Dataset for Evaluating Contextualized Representation of Biomedical Concepts in Language Models. **Scientific Data**, Vol. 11, No. 1, p. 455, 2024.

[9] George A. Miller. WordNet: A Lexical Database for English. In **Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994**, 1994.

[10] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. **Nucleic Acids Research**, Vol. 32, pp. D267–D270, 2004.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proc. of NAACL**, pp. 4171–4186, 2019.

[12] Kyoko Ohara. Relating Frames and Constructions in Japanese FrameNet. In **Proc. of LREC**, pp. 2474–2477, 2014.

[13] 永井宥之, 西山智弘, 大槻優佳, 藤牧貴子, 川端京子, 工藤紀子, 山崎由佳, 白石暖哉, 梶原智之, 進藤裕之, 河添悦. JMED-DICT: 大規模医療用語辞書の構築. 言語処理学会第 31 回年次大会, pp. 3509–3514, 2025.

[14] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In **Proc. of EMNLP**, pp. 230–237, 2004.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Joshi. Mandar, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. **arXiv:1907.11692**, 2019.

[16] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In **Proc. of ACL**, pp. 2526–2547, 2025.

[17] 杉本海人, 壹岐太一, 知田悠生, 金沢輝一, 相澤彰子. JMedRoBERTa: 日本語の医学論文にもとづいた事前学習済み言語モデルの構築と評価. 言語処理学会第 29 回年次大会, 2023.

[18] Koki Horiguchi, Tomoyuki Kajiwara, Takashi Ninomiya, Shoko Wakamiya, and Eiji Aramaki. MultiMSD: A Corpus for Multilingual Medical Text Simplification from Online Medical References. In **Findings of ACL**, pp. 9248–9258, 2025.

[19] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In **Findings of EMNLP**, pp. 1627–1643, 2020.

[20] Wataru Kawakami, Keita Suzuki, and Junichiro Iwasawa. Stabilizing Reasoning in Medical LLMs with Continued Pretraining and Reasoning Preference Optimization. **arXiv:2504.18080**, 2025.

[21] 山内崇史, 梶原智之, 荒瀬由紀. 文脈を考慮した単語ベクトル集合からの単語領域表現. 言語処理学会第 26 回年次大会, 2020.

[22] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In **Proc. of KDD**, p. 226–231, 1996.

[23] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In **Proc. of CoLM**, 2024.

[24] LLM-jp. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. **arXiv:2407.03963**, 2024.

[25] Qwen Team. Qwen2.5 Technical Report. **arXiv:2412.15115**, 2025.