JMedWiC: A Japanese Word-in-Context Dataset in the Medical Domain

Koki Horiguchi[1] **Seiji Sugiyama**[1] **Tomoyuki Kajiwara**[1,2]
Shoko Wakamiya[3] **Eiji Aramaki**[3]
Ehime University[1] The University of Osaka[2] Nara Institute of Science and Technology[3]
{horiguchi@ai., sugiyama@ai., kajiwara@}cs.ehime-u.ac.jp {wakamiya, aramaki}@is.naist.jp

Abstract

We release JMedWiC, a Japanese dataset for Word-in-Context (WiC) tasks specifically tailored to the medical domain. To address the challenge of word sense disambiguation, where the meaning of a word varies depending on its context, previous research has developed WiC datasets to evaluate word sense identity by determining whether a target word shares the same sense across two given contexts. In the medical domain, the misinterpretation of word senses can hinder the accurate comprehension of medical information; however, there is currently no Japanese WiC dataset specialized for this domain. Moreover, existing WiC datasets have been constructed using lexical resources with sense inventories, such as WordNet and UMLS, but such resources are not sufficiently developed for Japanese. Therefore, we construct a Japanese WiC dataset in the medical domain by manually annotating sense-identity labels for target words in context pairs automatically extracted from a large-scale corpus, without relying on lexical resources.

Keywords: Word-in-Context, Medical NLP, Japanese Dataset

1. Introduction

Word sense ambiguity, where a word takes on different meanings depending on the context, is one of the causes of performance degradation in downstream tasks such as machine translation (Chan et al., 2007) and information retrieval (Stokoe et al., 2003). To address this issue, research on word sense disambiguation (Navigli, 2009), which assigns appropriate sense labels to words in a context, has been conducted. In recent years, as a framework that does not rely on a sense label inventory, the Word-in-Context (WiC) task (Pilehvar and Camacho-Collados, 2019) has been proposed, which determines whether a target word has the same meaning in two different contexts. WiC datasets have been developed primarily for English (Breit et al., 2021; Pilehvar and Camacho-Collados, 2019; Raganato et al., 2020).

In the medical domain, many words share the same spelling as common words but have different meanings in specialized contexts. For example, the word "pocket" refers to a bag-like part sewn into clothing in everyday conversation. In surgery, it denotes a wound cavity larger than a skin defect, and in dentistry, it refers to the groove formed between a tooth and the surrounding gum. Such polysemy specific to the medical domain is not fully covered by existing WiC datasets. Therefore, while a WiC dataset specific to the medical domain has been constructed for English (Rouhizadeh et al., 2024), no such dataset exists for Japanese or other languages.

Furthermore, many existing WiC datasets are constructed based on lexical resources that explicitly define sense inventories, such as WordNet (Miller, 1994) and UMLS (Bodenreider, 2004), by extracting context pairs in which the target word has the same sense and context pairs in which it does not. In contrast, Japanese sense inventories specific to the medical domain are not sufficiently developed, and it is difficult to construct datasets using a similar approach.

In this study, we leverage contextualized word representations generated by BERT (Devlin et al., 2019) to construct a Japanese WiC dataset for the medical domain. Contextualized word representations are capable of capturing semantic differences of the same word based on its context as continuous representations. Leveraging this property, we automatically extract synonymous pairs and non-synonymous pairs and then perform manual annotation to construct JMedWiC[1], a Japanese WiC dataset for the medical domain. Our JMedWiC consists of two subsets, medical and general, each containing 1,000 context pairs.

2. Related Work

WiC (Pilehvar and Camacho-Collados, 2019) is a task that determines whether a target word has the same meaning in two different contexts, and datasets have been developed primarily for English (Breit et al., 2021; Pilehvar and Camacho-Collados, 2019; Raganato et al., 2020). The WiC datasets are centered on the lexical resource WordNet (Miller, 1994), which organizes words by sense units (synsets), and are constructed based on whether the target words share the same sense ID. In the medical domain, BioWiC (Rouhizadeh et al.,

[1] https://github.com/EhimeNLP/JMedWiC

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 222–227
16 May 2026.

	Target	Contexts (A / B)	Label
medical	痛み	**A:** この頃より肘と右肩の**痛み**に悩まされ始める。 (From around this time, he began to suffer from **pain** in his elbow and right shoulder.) **B:** エナメル質のう蝕では**痛み**を感じることはない。 (**Pain** is not felt in dental caries of the enamel.)	True
	胸壁	**A:** **胸壁**の良性腫瘍として最も頻度の高いのはどれか。 (Which is the most common benign tumor of the **chest wall**?) **B:** 銀地に、赤い 2 本の**胸壁**のある塔のある城。 (On a silver field, a castle with a tower featuring two red **breastworks**.)	False
general	四季	**A:** **四季**を通じての自然鑑賞地ともなっている。 (It is also a place for appreciating nature throughout the **four seasons**.) **B:** 大陸性気候に属する気候は比較的穏やかで、**四季**もある。 (Climates classified as continental are relatively temperate and feature **four seasons**.)	True
	足取り	**A:** PDF の初期の普及の**足取り**は緩やかなものであった。 (The early **adoption** of PDF was gradual.) **B:** 得意手は右四つ、押し、**足取り**。 (His preferred techniques are a right-hand inside grip, pushing, and **leg picks**.)	False

Table 1: Examples from JMedWiC

2024) has been constructed based on the Unified Medical Language System (UMLS) (Bodenreider, 2004), which integrates medical and biomedical concept systems, and is built according to whether the target word shares the same concept ID.

All of these existing datasets rely on lexical resources that explicitly maintain sense inventories, and in domains where such resources are unavailable, constructing datasets using the same methodology is challenging. In the Japanese medical domain, lexical resources such as WordNet or UMLS have not been sufficiently developed, and consequently, no WiC dataset has been constructed.

3. JMedWiC Datasets

To evaluate the ability to determine word sense identity in Japanese, we constructed JMedWiC, a WiC-format dataset. JMedWiC consists of two subsets: `medical`, which targets medical terms, and `general`, which targets general-domain terms. In this study, we select the target vocabulary for evaluation in each subset and collect their contextual occurrences from two different text corpora. We then automatically extract context pairs based on the semantic similarity of the target word and determine the final labels through manual annotation. Table 1 presents examples from our JMedWiC.

3.1. Target Vocabulary and Contexts

For the `medical` subset, we use JMED-DICT mini[2] as a vocabulary list for the medical domain. JMED-DICT is a dictionary that systematically collects vocabulary in the Japanese medical domain and consists of three sub-datasets: BODY, MEDICATION, and DISEASE. From this dictionary, we extracted 9,257 terms based on frequency information, and selected 6,476 terms consisting of 5 characters or fewer as target words. For the `general` subset, we extracted 7,500 content words from the high-frequency entries of the BCCWJ vocabulary list[3], and selected 7,395 words consisting of 5 characters or fewer as target words.

In this study, we use Wikipedia[4] as a corpus covering a wide range of general-domain contexts. In addition, to ensure the inclusion of medical-domain contexts, we collected medical texts from the online medical encyclopedia, the MSD Manuals[5]. For each corpus, we applied a rule-based sentence segmentation method[6]. For the `medical` subset, sentences containing the target words were extracted from both Wikipedia and the MSD Manuals, whereas for the `general` subset, such sentences were extracted from Wikipedia. To accurately identify the occurrence positions of the target words, we applied word segmentation using MeCab (Kudo et al., 2004)[7], incorporating a medical dictionary[8]. In addition, to exclude extremely short or overly long

[2] https://sip3-d2.naist.jp/jmed-dict.html

[3] Short-unit word list ver.1.0 of the Balanced Corpus of Contemporary Written Japanese (BCCWJ).

[4] https://huggingface.co/datasets/range3/wiki40b-ja

[5] https://www.msdmanuals.com

[6] https://github.com/wwwcojp/ja_sentence_segmenter

[7] https://taku910.github.io/mecab/

[8] https://sociocom.naist.jp/j-meddic-for-mecab/

sentences, only sentences with lengths between 10 and 50 characters were retained.

3.2. Context Pairing

To efficiently extract candidate pairs for manual annotation, we propose a context-pairing method based on the cosine similarity of contextualized word embeddings. Since contextualized word embeddings represent the same word differently depending on its context, the semantic closeness of a word can be estimated based on the cosine similarity between its vectors in two different contexts.

In this method, each context is fed into a masked language model, and the hidden-state vector corresponding to the target word token is used as the contextualized word embedding. If the target word is segmented into multiple subwords, the mean of their vectors is used. The cosine similarity between these vectors is computed for context pairs containing the same target word. We extracted pseudo-synonymous pairs and pseudo-non-synonymous pairs using the predetermined thresholds θ_{high} and θ_{low}.

3.3. Automatic Construction of Dataset

To select a masked language model suitable for context pairing, we employed a word alignment task to measure the models' ability to identify semantic correspondences between words. Four Japanese masked language models were evaluated: BERT[9] (Devlin et al., 2019), RoBERTa[10] (Liu et al., 2019), ModernBERT[11] (Warner et al., 2025), and JMedRoBERTa[12]. For evaluation, we used 300 sentence pairs extracted from the Japanese evaluation set of the medical text simplification parallel corpus[13] (Horiguchi et al., 2025), which consists of semantically aligned complex and simplified medical sentences. For each sentence, word segmentation was performed using MeCab (Kudo et al., 2004) augmented with the medical dictionary[7], and the first author manually annotated the word alignments. For the alignment method, we used SimAlign (Jalili Sabet et al., 2020), which pairs word tokens that maximize the cosine similarity of their contextualized embeddings[14], and performance was evaluated using the F-score.

[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

[13] https://github.com/EhimeNLP/MultiMSDcorpus

[14] SimAlign parameters: bpe, argmax, distortions=0.1

	Precision	Recall	F-score
BERT	65.51	76.64	**70.64**
RoBERTa	65.63	72.08	68.70
ModernBERT	64.31	76.71	69.96
JMedRoBERTa	53.30	77.76	63.24

Table 2: Results of Word Alignment Using SimAlign

	True	False	Total
medical	570 (700)	430 (300)	1,000
general	597 (500)	403 (500)	1,000

Table 3: Label distribution of JMedWiC (parentheses show automatic extraction counts).

As shown in Table 2, BERT achieved the highest F-score. Therefore, in this study, we adopt BERT and perform the context pairing described in Section 3.2 on the set of contexts extracted in Section 3.1. To avoid pairing non-medical contexts within the `medical` subset, we imposed the condition that at least one context in each pair must be derived from the MSD Manuals. In context pairing, pairs with a cosine similarity of $0.75 \leq \theta_{\text{high}} \leq 0.80$ were classified as pseudo-synonymous pairs, while those with a cosine similarity less than $\theta_{\text{low}} = 0.6$ were classified as pseudo-non-synonymous pairs. From the candidate set obtained through context pairing, we selected a total of 1,000 pairs for the `medical` subset, consisting of 700 pseudo-synonymous pairs and 300 pseudo-non-synonymous pairs, and a total of 1,000 pairs for the `general` subset, consisting of 500 pseudo-synonymous pairs and 500 pseudo-non-synonymous pairs. The lower proportion of non-synonymous pairs in the `medical` subset is due to the limited size of the MSD Manuals compared to Wikipedia, which restricted the number of contexts available for extraction. Since the contexts collected from the MSD Manuals are subject to redistribution restrictions, the corresponding contexts among the selected 1,000 pairs were rephrased while preserving their meaning using MedQwen-72b[15] (Kawakami et al., 2025), a large language model specialized in the medical domain.

3.4. Synonymy Annotation

For the collected context pairs, one medical professional annotated the `medical` subset, and one student annotated the `general` subset, indicating whether the target word had the same meaning in each pair. As a result of the annotation, WiC-format datasets consisting of 1,000 pairs each were con-

[15] https://huggingface.co/pfnet/Preferred-MedLLM-Qwen-72B

		medical			general		
		Precision	Recall	F-score	Precision	Recall	F-score
Masked Language Model	BERT	0.693	0.868	0.771	0.839	0.760	0.798
	RoBERTa	0.658	0.854	0.744	0.598	0.980	0.743
	ModernBERT	0.628	0.875	0.732	0.598	0.993	0.746
	JMedRoBERTa	0.567	0.956	0.711	0.598	0.997	0.747
Masked Language Model + Clustering	BERT	0.650	0.840	0.733	0.660	0.660	0.660
	RoBERTa	0.572	0.937	0.710	0.574	0.895	0.699
	ModernBERT	0.570	0.984	0.722	0.569	0.993	0.723
	JMedRoBERTa	0.530	0.751	0.621	0.569	0.970	0.717
Large Language Model	Swallow-8b	0.838	0.570	0.678	0.860	0.216	0.345
	llmjp-13b	0.853	0.409	0.553	0.886	0.131	0.228
	Qwen-72b	0.885	0.637	0.741	0.929	0.305	0.459
	MedQwen-72b	0.882	0.563	0.687	0.889	0.362	0.514

Table 4: Evaluation results of word sense identity determination in `medical` and `general` domains.

structed for both subsets (Table 3). The consistency between automatically assigned labels, determined using the cosine similarity thresholds, and the manual annotations was 81.2% for the `medical` subset and 82.3% for the `general` subset.

4. Experiments

We evaluate the performance of three unsupervised methods on JMedWiC.

4.1. Experimental Setup

Masked Language Model Each context is input to BERT, and the subword embeddings corresponding to the target word are averaged to obtain a contextualized word embedding. The cosine similarity between the word embeddings obtained from the two contexts is computed. If the similarity exceeds the threshold θ, the meanings are considered identical; if it is below θ, the meanings are considered different. We used four Japanese BERT models: BERT[9], RoBERTa[10], ModernBERT[11], and JMedRoBERTa[12]. The threshold θ was selected from $\{0.50, 0.55, \ldots, 0.95\}$ as the value that maximized the F-score. It should be noted that BERT was used during dataset construction (Section 3.3), which gives it a favorable condition compared with the other models.

Masked Language Model + Clustering We employ DBSCAN (Ester et al., 1996), a density-based clustering method that does not require a predetermined number of clusters, to evaluate word sense identity. All contexts containing the target words were collected from Wikipedia[4] and the MSD Manuals[5] for the `medical` subset and from Wikipedia for the `general` subset, and contextualized word embeddings were obtained for these contexts. DBSCAN is then applied to the set of contextualized embeddings, including the contexts to be evaluated, and two contexts are considered to have the same meaning if they belong to the same cluster, and different meanings if they belong to different clusters. The same four Japanese masked language models as in the previous section were used.

Large Language Model In zero-shot in-context learning with a large language model (LLM), the model is provided with an instruction, the target word, and a context pair, and is prompted to output whether the target word has the "same" or "different" meaning in the two contexts. The models used include the general-purpose LLMs Swallow-8b[16] (Fujii et al., 2024), llmjp-13b[17] (LLM-jp, 2024), Qwen-72b[18] (Qwen Team, 2025), and the medical-domain LLM MedQwen-72b[15] (Kawakami et al., 2025). The prompt given to the LLMs is as follows:

Prompt

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。(Please determine whether the target word has the same meaning in the following two sentences. If the meaning is the same, answer "same"; if the meaning is different, answer "different.")

4.2. Results

The experimental results are presented in Table 4. BERT exhibited markedly high performance in

[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

both domains, as the cosine similarity was computed based on the same contextualized embeddings used during dataset construction. Excluding BERT, RoBERTa with a cosine similarity threshold achieved the highest performance on the `medical` evaluation, while those of JMedRoBERTa achieved the best performance on the `general` evaluation. Overall, masked language model + clustering demonstrated comparable performance compared with masked language models using optimal cosine similarity thresholds, whereas zero-shot in-context learning with LLM demonstrated a decrease in performance.

When comparing across domains, performance in the `medical` subset tended to surpass that in the `general` subset. Medical terms are often used in specialized contexts, and their range of meanings is relatively limited, which may have made sense disambiguation easier. In contrast, general terms often appear in contexts that are polysemous or figurative, making sense boundaries more ambiguous and thereby increasing the difficulty of sense disambiguation.

Comparing general purpose and medical specialized models on the `medical` subset, no clear advantage was observed for the medical-specialized models; both JMedRoBERTa and MedQwen-72b performed worse than the general-purpose models. This result suggests that what is important for the WiC task may not be domain-specific knowledge about the target words themselves, but rather general linguistic knowledge for identifying sense equivalence from context.

5. Conclusion

In this study, we constructed JMedWiC, consisting of 1,000 pairs each for the medical and general domains, to evaluate the ability to determine context-dependent word sense identity in Japanese medical texts, using contextualized word embeddings for context pairing and manual annotation. Experiments showed that masked language models based on cosine similarity thresholds performed best in both the medical and general domains, outperforming both masked language model + clustering and large language models. In addition, it was revealed that medical-specialized models do not necessarily outperform general-purpose models, and that the task difficulty differs across domains.

Acknowledgements

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on “Integrated Health Care System” Grant Number JPJ012425.

Bibliographical References

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

Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. WiC-TSV: An Evaluation Benchmark for Target Sense Verification of Words in Context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1635–1645.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page 226 – 231.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In *Proceedings of the First Conference on Language Modeling*.

Koki Horiguchi, Tomoyuki Kajiwara, Takashi Ninomiya, Shoko Wakamiya, and Eiji Aramaki. 2025. MultiMSD: A Corpus for Multilingual Medical Text Simplification from Online Medical References. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9248–9258.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643.

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

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

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

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

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

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1–10:69.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273.

Qwen Team. 2025. Qwen2.5 Technical Report. *arXiv:2412.15115*.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7193–7206.

Hossein Rouhizadeh, Irina Nikishina, Anthony Yazdani, Alban Bornet, Boya Zhang, Julien Ehrsam, Christophe Gaudet-Blavignac, Nona Naderi, and Douglas Teodoro. 2024. A Dataset for Evaluating Contextualized Representation of Biomedical Concepts in Language Models. *Scientific Data*, 11(1):455.

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

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. 2025. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.