Detecting *Honkadori* based on *Waka* Embeddings

Hayato Ogawa, Kaito Horio, Daisuke Kawahara
Waseda University
{cookie3120@ruri.,kakakakakakaito@akane.,dkw@}waseda.jp

Abstract

We develop an embedding model specifically designed for Waka poetry and use it to build a model for detecting *Honkadori*. *Waka* is a traditional form of old Japanese poetry that has been composed since ancient times. *Honkadori* is a sophisticated poetic technique in Japanese classical literature where poets incorporate words or poetic sentiments from old *Waka*s (*Honka*) into their own work. First, we fine-tune a pre-trained language model using contrastive learning to construct a *Waka*-specialized embedding model. Then, using the embedding vectors obtained from this model and features extracted from them, we train a machine learning model to detect the *Honka* (original poem) of *Waka*s that employ the *Honkadori* technique. Using paired data of *Honka* and *Waka*s that are considered to use *Honkadori*, we evaluated the *Honka* detection model and demonstrated that it can detect *Honka* with reasonable accuracy.

1 Introduction

Waka is a traditional form of Japanese poetry based on combinations of 5- and 7-syllable units. First appearing in the *Nara* period (early 8th to late 8th century), *Waka* continued to be composed through the *Edo* period (early 17th to late 19th century). *Waka* can take various forms, such as repetitions of 5-7-5 syllable patterns or a 38-syllable structure (5-7-7-5-7-7), but the most common form is the 31-syllable structure (5-7-5-7-7), known as *Tanka*. Since its inception in the *Nara* period, people have gathered to compose and recite *Waka* on common themes, and it became so deeply rooted in Japanese culture that emperors would sometimes order the compilation of *Waka* collections.

Honkadori is a sophisticated poetic technique in Japanese classical literature for composing *Waka*, where poets incorporate words or poetic sentiments from old *Waka*s (*Honka*) into their own work.

Figure 1: Example of *Honkadori*. The upper poem shows the HONKA: "How painful it is, the rain falling at Miwa promontory—at the riverbank of Sano (佐野の渡り), there is not even a house to shelter in." The lower poem shows the HONKADORI: "At dusk, I stopped my horse at the riverbank of Sano (佐野の渡り), brushing off the snow from my sleeves, yet finding no shelter anywhere." The HONKADORI alludes to the phrase "佐野の渡り" (the riverbank of Sano) from the HONKA, sharing the common theme that there is nowhere to hide from the weather.

This technique creates layered meanings while expressing their own poetic voice (Ooka, 2009). In the practice of *Honkadori*, the original *Waka* that serves as the source of borrowed words or expressions is called *Honka*. This technique differs from simple quotation or plagiarism, as it requires deep understanding and creative reinterpretation of classical works. Interestingly, a similar practice exists in modern music, particularly in hip-hop, called sampling. Sampling is a music production technique where parts of existing songs (such as drums,

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 220–227

bass, melody, or vocals) are extracted and reconstructed within new compositions. For example, Kanye West's "Gold Digger" is known for sampling Ray Charles's "I Got a Woman," adding new interpretations to the original work. Like *Honkadori* in *Waka*, sampling represents a creative technique where modern musicians show respect for classic works while adding their own interpretations. An example of *Honkadori* is shown in Figure 1.

Honkadori is said to have been established during the *Heian* period (late 8th to late 12th century). *Teika Fujiwara*, a prominent poet from the *Heian* period, established the following rules for *Honkadori* in his poetic treatise *Eikataigai* (The Editorial Committee of the Great Dictionary of Waka Literature, 2014):

- One should not borrow from *Waka*s of contemporary poets.
- The borrowed phrases from classical *Waka*s should be limited to approximately two phrases.
- The theme must be different from the *Honka*.

We propose a method for automatically detect *Honkadori* in *Waka*. While shared characters between *Waka*s provide important clues for *Honkadori* detection, the second rule from *Eikataigai* often results in relatively short common subsequences between the *Honka* and the *Waka* employing *Honkadori* (hereafter, this is denoted as HONKADORI to distinguish it from the technique itself, and the original *Waka* that serves as the source of this HONKADORI is denoted as HONKA). Therefore, character-based methods alone struggle to automatically distinguish *Honkadori* from other similar *Waka*s. To address this problem, we first develop a *Waka*-specialized embedding model and then create a model that calculates the probability of any given pair of *Waka*s being in a *Honkadori* relationship. Our study is expected to contribute to classical literature studies through the detection of previously undiscovered instances of *Honkadori*.

2 Related Work

2.1 Character-based Similar *Waka* Detection Methods

Yamazaki et al. (1998) and Takeda et al. (2000) have proposed methods for detecting similar *Waka*s based on character similarity. These methods enable the detection of various types of similar *Waka*s, including *Honkadori*, expressions used in specific poetic situations, variant *Waka*s that developed different expressions through transmission, and *Waka*s that share rhetorical devices such as *makurakotoba* (set epithets in classical Japanese poetry). However, these studies do not focus on the semantic aspects of *Waka*s, making it difficult to detect pairs of similar *Waka*s that do not share significant character similarities.

2.2 Allusion Detection Methods Using Embedding Vectors

Kondo (2024) has proposed a method for detecting *Hikiuta* (poetic allusions) using embedding vectors. In their study, they focused on identifying allusions between two significant classical Japanese works that are *The Tale of Genji* and *Kokin Wakashū*. *The Tale of Genji* is a long narrative work, or novel, written by Murasaki Shikibu during the middle Heian period. The *Kokin Wakashū* is a *Waka*'s anthology compiled in the early Heian period under the Imperial command of the Emperor at that time. *The Tale of Genji* contains several passages that use *Hikiuta* based on *Waka* poems included in the *Kokin Wakashū*. *Hikiuta* is a technique similar to *Honkadori*, where a famous *Waka* passage is quoted within prose text or an emotional passage (Nishizawa, 2002).

To detect such *Hikiuta*, Kondo (2024) has proposed a method using embedding vectors. This method first embeds text segments from *The Tale of Genji* and *Kokin Wakashū* into a vector space using OpenAI's text-embedding-ada-002 model (OpenAI, 2022). Then, it calculates cosine similarities between the embedding vector of each *Waka* from *Kokin Wakashū* and the embedding vectors of text segments from *The Tale of Genji* and identifies high-similarity pairs as potential allusions. Furthermore, the study reports that applying N-gram character matching as a filter increases the proportion of verifiable allusions among the candidates in *The Tale of Genji*. This method has also led to the discovery of previously unidentified allusions. However, the study does not evaluate either the accuracy of classical text embeddings of text-embedding-ada-002 or the precision of the allusion detection method itself.

3 Construction and Evaluation of *Waka* Embedding Models

We develop *Waka* embedding models. We first construct a training dataset and then use it to fine-tune a pre-trained encoder model with unsupervised SimCSE (Gao et al., 2021). We evaluate the resulting models using pairs of *Waka*s from *Hyakunin Isshu* (en: One Hundred *Waka*s by One Hundred Poets) and their modern Japanese translations.

3.1 Construction of Training Datasets

To train *Waka* embedding models, we use literary works from the *Nara* period through the *Edo* period recorded in the Corpus of Historical Japanese (CHJ) (NINJAL, 2024). Additionally, we use *Tanka*s from the Modern *Tanka* Database (Yuna et al., 2022) and literary works in classical Japanese orthography published in the *Aozora Bunko* digital library.

From CHJ, we obtained approximately 100,000 sentences including approximately 17,000 *Waka*s (referred to as the **CHJ dataset**), approximately 140,000 *Waka*s from the Modern *Tanka* Database (referred to as the **Modern *Tanka* dataset**), and approximately 335,000 sentences from *Aozora Bunko* (referred to as the ***Aozora* dataset**).

3.2 Construction of *Waka* Embedding Models

In supervised learning for text embedding models, we need annotations indicating which sentences are semantically similar and which are different. However, creating such annotations for large amounts of text is time-consuming and costly. To avoid the annotation cost, we fine-tune a Japanese RoBERTa model (Liu et al., 2019)[1] using unsupervised SimCSE, a contrastive learning approach. Unsupervised SimCSE generates two slightly different embedding vectors by applying dropout twice to the same sentence and treats these as positive examples. This approach allows us to train effective embedding models without the need for manual annotation. When inputting text into this model, we perform word segmentation using the Juman++ morphological analyzer (Tolmachev et al., 2018). We compare the performance of unsupervised SimCSE using individual datasets constructed in Section 3.1 and combined datasets.

[1]https://huggingface.co/nlp-waseda/roberta-base-japanese

3.2.1 Models with Individual Datasets

We trained a model for 5 epochs using each of the CHJ dataset, Modern *Tanka* dataset, and *Aozora* dataset individually.

3.2.2 Models with Combined Datasets

We trained a model for 5 epochs using a dataset created by merging and shuffling the *Aozora*, Modern *Tanka*, and CHJ datasets. Furthermore, we implemented curriculum learning (Bengio et al., 2009) that gradually adapts the training data to the *Waka* format as follows. In this curriculum learning process, datasets other than the CHJ dataset were used for only 1 epoch of training, followed by 5 epochs of training with the CHJ dataset.

- *Aozora dataset* → CHJ dataset
- Modern *Tanka* dataset → CHJ dataset
- *Aozora* dataset → Modern *Tanka* dataset → CHJ dataset

3.3 Evaluation of *Waka* Embedding Models

3.3.1 Evaluation Method

To quantitatively evaluate the performance of the trained *Waka* embedding models, we adopt an evaluation method using a parallel corpus of all 100 *Waka*s from *Hyakunin Isshu* and their modern Japanese translations. *Hyakunin Isshu* is an anthology of 100 *Waka*s, with a *Waka* carefully selected to represent each of one hundred distinct poets. The parallel corpus was obtained from the website "History of Hyakunin Isshu"[2] [3] [4]. The evaluation was conducted according to the following procedure:

1. Convert an original *Waka* into an embedding vector using the target model.
2. Similarly, convert each of the 100 modern translations into an embedding vector using the same model.
3. Calculate cosine similarities between the embedding vectors of the original *Waka* and each of all modern translations.
4. For the original *Waka*, consider the modern translation with the highest similarity as the model's predicted translation.
5. Count a prediction as correct if the predicted translation matches the true translation and evaluate the model using accuracy over all 100 *Waka*s.

[2]https://hyakunin.stardust31.com/gendaiyaku.html
[3]https://hyakunin.stardust31.com/gendaiyaku-itiran.html
[4]https://hyakunin.stardust31.com/yaku.html

Model	text-embedding-small	text-embedding-large	text-embedding-ada-002	*Waka* embedding model
Accuracy	0.95	0.91	0.92	0.95

Table 1: Accuracy comparison between OpenAI models and *Waka* embedding model.

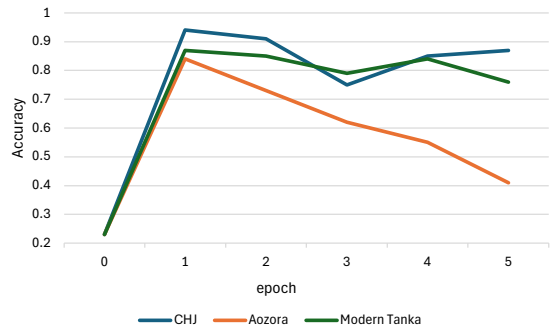

Figure 2: Accuracy transitions by epoch for training with individual datasets.

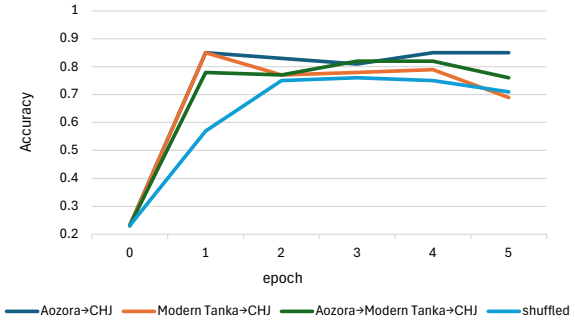

Figure 3: Accuracy transitions by epoch for training with combined datasets.

3.3.2 Evaluation Results

Figure 2 shows the accuracy transitions for each epoch in training with the individual datasets. The best-performing model using a single dataset was the one trained for 1 epoch on the CHJ dataset, achieving an accuracy of 0.95. Figure 3 shows the accuracy transitions for each epoch in training with the combined datasets. Multiple models achieved the highest performance with a combined dataset, with an accuracy of 0.85. Therefore, the model trained for 1 epoch on the CHJ dataset demonstrated the highest performance. Based on these results, we adopted the model trained for 1 epoch on the CHJ dataset as our *Waka* embedding model and used it in the subsequent experiments.

3.3.3 Comparison with OpenAI Models

We compared OpenAI's text embedding models with our *Waka* embedding model using the evaluation method described in Section 3.3.2. The results are shown in Table 1. Among the OpenAI models, text-embedding-3-small achieved the highest performance with an accuracy of 0.95. Our *Waka* embedding model demonstrated performance equivalent to it.

4 Construction of HONKA Detection Models

To automatically detect the *Honkadori* technique, we need to consider not only surface similarities between *Waka*s but also their semantic relationships. Therefore, we construct a HONKA detection model that uses machine learning to understand the relationship between HONKA and their HONKADORI by using features obtained from our *Waka* embedding model. The construction of this model involves three steps: first collecting training data of *Honkadori* pairs, then training machine learning models using features extracted from *Waka* pairs, and finally evaluating the model's performance.

4.1 Dataset Construction

To construct our training and evaluation datasets for the *Honka* detection model, we collected positive examples and two distinct types of negative examples.

First, as positive examples, we manually collected 300 pairs of HONKA and their corresponding HONKADORI from the Eight Imperial Anthologies[5] as documented in the 日本うたことば表現辞典本歌本節取編 (en: Dictionary of Japanese Poetic Expressions - Compilation of Honka and Honsetsudori). From these collected pairs, we allocated 200 pairs for the training dataset and the remaining 100 pairs for the evaluation dataset.

We created two distinct sets of negative examples. First, we constructed a dataset of 200 randomly combined *Waka* pairs from the Eight Imperial Anthologies. These pairs serve as our first type of negative examples, representing arbitrary combinations of poems without any intentional relationship. For our second set of negative examples, we focused on poems sharing *makurakotoba*, i.e., fixed epithetic expressions that precede and modify

[5]The Eight Imperial Anthologies (*Hachidaishū*) are the most prestigious collections of *Waka*, compiled by imperial order.

specific words through conventional associations. We collected these pairs from the Dictionary of Japanese Poetic Expressions: Makurakotoba Volume 1 and 2 (日本うたことば表現辞典枕詞編(上・下)) (Ooka, 2007). To manage the collection process efficiently, we selected 10 types of *makurakotoba* and collected 6 *Waka*s for each type. We then created all possible unordered pairs from each set of 6 *Waka*s, which resulted in 15 pairs per *makurakotoba* type. This process yielded a total of 150 pairs of *Waka*s that share *makurakotoba* but are not classified as HONKADORI.

In summary, we constructed a dataset with 300 pairs of *Honka* and their corresponding *Honkadori* (200 pairs for training and 100 pairs for evaluation) as positive examples. We also constructed 350 pairs of *Waka*s as negative examples, including 150 pairs that serve as hard negative examples. All negative examples are used only for training purposes.

4.2 Machine Learning Model Construction

We constructed the HONKA detection model based on embedding vectors obtained from the *Waka* embedding model. The HONKA detection model calculates the probability that a pair of *Waka* is in a *Honkadori* relationship based on features extracted from the pair. Our *Waka* embeddings (RoBERTa base) are 768-dimensional, meaning that using embedding vectors for both *Waka*s in a pair would result in a 768×2 dimensional input. Due to the limited amount of training data, we intended to restrict the input dimensionality of the machine learning model. Therefore, instead of using embedding vectors directly, we used the following seven features:

- Cosine similarity between *Waka* pairs
- Top 5 highest cosine similarities from the 25 similarities between corresponding phrases (5-7-5-7-7) of the *Waka* pairs
- Longest common subsequence length between *Waka* pairs

To minimize the impact of orthographic variations, the longest common subsequence length between *Waka* pairs is calculated by obtaining readings using the morphological analyzer MeCab (Kudo et al., 2004). These readings are obtained with the *Waka*-specific morphological dictionary *Waka* UniDic (Ogiso et al., 2012). The readings are then converted to *kana* characters with voiced and semi-voiced sound marks removed. Using these features,

Method	1st	2nd	3rd	4th	5th
Nearest Neighbor	8	9	2	0	1
Logistic Regression	8	9	2	0	1
SVM	7	0	0	0	1
LightGBM	1	6	2	1	2
MLP	9	11	1	2	0
Meta-model	10	5	0	0	0

Table 2: Rank distribution of HONKA detection results of each method.

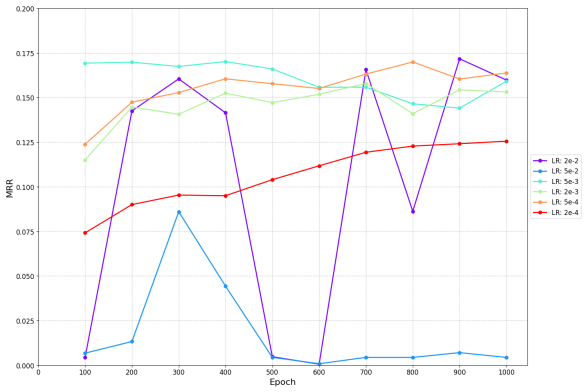

Figure 4: MRR transitions for different learning rates in MLP.

we trained logistic regression, SVM, LightGBM, MLP models, and a meta-model (logistic regression) that blends these models. For MLP, we conducted training with multiple learning rates. The detailed training settings are provided in Table 6.

4.3 Experiments

4.3.1 Evaluation Method and Baseline

Evaluation Method We evaluated the accuracy of HONKA detection using approximately 9,600 *Waka* from the Eight Imperial Anthologies in the Corpus of Historical Japanese (referred to as the **Eight Imperial Anthologies Dataset**) and the *Honkadori* evaluation dataset (100 pairs). The evaluation was conducted according to the following procedure:

1. Apply the HONKA detection method to each HONKADORI in the evaluation dataset and all *Waka*s in the Eight Imperial Anthologies Dataset.
2. Sort the Eight Imperial Anthologies Dataset based on the probabilities output by the model in descending order of HONKA likelihood.
3. Evaluate using the following two metrics:

Nearest Neighbor	Logistic Regression	SVM	LightGBM	MLP	Meta-model
0.149	0.147	0.0793	0.0650	0.172	0.137

Table 3: MRR for each HONKA detection method.

HONKADORI	九重の (imperial court) にほひなりせば (if still as precious) さくらばな (cherry blossoms) 春知りそむる (just learning spring) かひやあらまし (would have had meaning) (en: If this place was still as precious as it was back then, these cherry blossoms would have held more meaning.)
Rank	Predicted HONKA
1 (HONKA)	ことしより (from this year onward) 春しりそむる (just learning spring) さくらばな (cherry blossoms) ちるといふことは (the act of scattering) ならはざらなん (please do not learn) (en: Please don't learn how to scatter, oh cherry blossoms that have just begun to bloom this year as if you've only just discovered spring.)
2	さくら花 (cherry blossoms) そこなる影ぞ (reflection there) おしまるる (is regrettable) しづめる人の (of the sad people) 春とおもへば (when I think of spring) (en: The cherry blossoms are blooming. When I see their reflection in the pond, it reminds me of those who are unhappy.)
3	さくら花 (cherry blossoms) 匂ふなごりに (in their lingering fragrance) 大かたの (all of) 春さへ (even spring) おしくおもほゆるかな (feels precious indeed) (en: In the lingering beauty of the cherry blossoms in full bloom, even the entire spring becomes precious, and I cannot help but feel this way.)

Table 4: Example of correct HONKA prediction ranked first by the model. The bold text represents the shared character sequences between HONKADORI and HONKA.

- Top-5 correct count: The number of times the correct HONKA appeared in the top 5 entries of the sorted Eight Imperial Anthologies Dataset.
- MRR (Mean Reciprocal Rank): The average of the reciprocal of the rank at which the correct HONKA appeared.

Baseline As a baseline for comparing our proposed method, we used HONKA detection based on nearest neighbor search. We calculated cosine similarities between the vectors of HONKADORI in the evaluation dataset and each of the *Waka*s in the Eight Imperial Anthologies Dataset. The evaluation was performed by sorting the Eight Imperial Anthologies Dataset in descending order of cosine similarity.

4.3.2 Experimental Results

Table 2 shows the rank distribution of HONKA detection results of each method. Table 3 shows the MRR of each model alongside the baseline MRR. The specifications of each model are shown in Table 6.

While logistic regression and MLP showed relatively good results, SVM and LightGBM performed significantly worse than the baseline. The model with the highest top-5 correct count was MLP (learning rate 2e-2, 700 epochs) with 23 cases. The model with the highest MRR was MLP (learning rate 2e-2, 900 epochs) with 0.172, indicating higher detection accuracy for HONKADORI than the nearest neighbor search. Table 4 shows an example where the model correctly identified the HONKA with the highest probability. Additional examples of *Honkadori* pairs that were included in the top-3 predictions by the model are shown in Table 5. In Table 5, "rank" refers to the position of each HONKA when sorted based on the probability output by the model in descending order of HONKA likelihood. Furthermore, Figure 4 shows the results of comparative experiments with different learning rates for MLP.

5 Conclusion

We constructed *Waka*-specialized embedding models and HONKA detection models. Furthermore, by building machine learning models using features extracted from the embedding vectors output by these models, we demonstrated that HONKA detection is possible with reasonable accuracy.

This study has several challenges to address. First, Juman++, which was used for input to the *Waka* embedding model, is a morphological analyzer designed for modern Japanese and is not well-suited for tokenizing old Japanese texts. Next is the amount of training data. While we manually collected 300 pairs of *Waka* with *Honkadori* relationships, higher accuracy could be expected with

Rank	HONKA	HONKADORI
1	あだなりと (vainly known) 名にこそたてれ (though bearing the name) 桜花 (cherry blossoms) としにまれなる (rarely each year) 人もまちけり (still I wait for someone) (en: Though cherry blossoms are known to scatter so easily, I still wait for those who visit but rarely in a year.)	嵐吹く (storm-blown) 花の梢は (the tips of blossoms) あだなりと (vainly known) 名にこそたてれ (though bearing the name) 花の白雲 (white clouds of flowers) (en: Though the storm-blown cherry blossoms are known to scatter easily, they still grace the sky like clouds of white flowers.)
2	けふこずは (if you don't come today) あすは雪とぞ (tomorrow surely snow) 降なましき (will fall like) えずは有とも (even if they remain) 花とみまし (would you see them as flowers?) (en: If you do not come today, these cherry blossoms will scatter and fall like snow. Unlike snow, even if they remained without fading, would they still be seen as flowers?)	さくら色の (cherry-colored) 庭の春風 (spring breeze in the garden) あともなし (no trace remains) 訪はばぞ人の (if someone were to visit) 雪とだにみん (might see them at least as snow) (en: The spring wind that once carried cherry blossom petals through my garden has left no trace behind; if only someone would visit, they might see the scattered petals as fallen snow and find beauty in the scene, but with no visitors, not even footprints remain.)
3	山たかみ (high in the mountains) 人もすさめぬ (ignored by people) 桜花 (cherry blossoms) いたくなわびそ (do not grieve so deeply) 我見はやさむ (I shall come to see you) (en: Cherry blossoms on the high mountain, though others pass you by without care, do not grieve so deeply—for I shall admire you and sing your praise.)	春くれど (though spring has come) 人もすさめぬ (ignored by people) 山桜 (mountain cherry blossoms) 風のたよりに (guided by the wind) 我のみぞとふ (only I come to visit) (en: Though spring has come, no one pays heed to the mountain cherry blossoms—only I, guided by the wind, go to visit them.)

Table 5: Examples of correctly predicted HONKADORI within Top-3 rankings.

access to more data. Additionally, while the *Waka* embedding model evaluation uses classical texts and their modern Japanese translations, it would be preferable to construct an evaluation dataset composed entirely of classical texts. By addressing these challenges, we can expect further improvements in model accuracy and greater contributions to classical literature research. Moreover, it is expected to have a wide range of applications not only in classical literature research but also in identifying text reuse in modern internet memes, literature, visual works, and other media.

Acknowledgments

We would like to express our sincere gratitude to Ms. Yuna Murata for providing the Modern Tanka Database for this research. This work was supported by JSPS KAKENHI Grant Numbers JP23K22374 and JP24H00727.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yasuhiro Kondo. 2024. Detecting implicitly quoted waka within 'the tale of genji' through vector search. *The 38th Annual Conference of the Japanese Society for Artificial Intelligence*.

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, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv. Abs/1907.11692.

NINJAL. 2024. The corpus of historical japanese.

Masashi Nishizawa. 2002. *Dictionary of Terms for Reading Japanese Classical Literature*. Tokyodo Shuppan.

Toshinobu Ogiso, Mamoru Komachi, Yasuharu Den, and Yuji Matsumoto. 2012. Unidic for early middle japanese: a dictionary for morphological analysis of classical japanese. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 911–915.

Mkoto Ooka. 2007. 日本うたことば表現辞典枕詞編*(上・下)(en: Dictionary 275 of Japanese Poetic Expressions: Makurakotoba Volume 1 and 2)*. yushikan.

Model	Parameter	Value
MLP	Input layer	Dimension: 7
	Hidden layer 1	Fully connected layer ($7 \rightarrow 32$), Activation function: ReLU, Dropout rate: 0.2
	Hidden layer 2	Fully connected layer ($32 \rightarrow 16$), Activation function: ReLU, Dropout rate: 0.2
	Output layer	Fully connected layer ($16 \rightarrow 1$), Activation function: Sigmoid
	Batch size	8
	Weight decay	1×10^{-3}
	Optimizer	Adam
	Loss function	Binary Cross Entropy
	Input preprocessing	StandardScaler
Logistic Regression	Maximum iterations	1000
	Regularization	L2
SVM	Kernel function	RBF
	Probability estimates	Enabled
	Regularization parameter (C)	1.0
LightGBM	Objective function	binary
	Evaluation metric	binary error
	Number of boosting rounds	100
Meta-model	Input features	First-layer prediction probabilities (Logistic Regression, SVM, LightGBM, MLP)
	Data split configuration	60% of training data used for first-layer learning 40% of training data used for meta-model learning
	Maximum iterations	1000
	Regularization	L2
	Tolerance for stopping criteria	1×10^{-4}

Table 6: Specifications of HONKA detection models.

Mkoto Ooka. 2009. 日本うたことば表現辞典本歌本節取編*(en: Dictionary of Japanese Poetic Expressions - Compilation of Honka and Honsetsudori)*. yushikan.

OpenAI. 2022. New and improved embedding model. Accessed: 2024-10-26.

Masayuki Takeda, Tomoko Fukuda, Ichiro Nanri, Mayumi Yamazaki, and Tamari Koichi. 2000. 和歌データからの類似歌発見(en: Discovery of similar waka from waka datas). 統計数理*(en: Statistical Mathematics)*, 48(2):289–310.

The Editorial Committee of the Great Dictionary of Waka Literature. 2014. 和歌文学大辞典*(en: the Great Dictionary of Waka Literature)*. Tokyodo Shuppan.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium. Association for Computational Linguistics.

Mayumi Yamazaki, Masayuki Takeda, Tomoko Fukuda, and Ichiro Nanri. 1998. 和歌データベースからの類似歌の自動抽出(en: Automatic extraction of similar poems from a waka database). 人文科学とコンピュータ*(en: Humanities and Computing)*, 40(8):40–48.

Murata Yuna, Kiyonori Nagasaki, and Ikki Ohmukai. 2022. 近代短歌全文テキストデータベースの構築(en: Building a comprehensive text database of modern tanka poetry). *Digital Humanities*, 3(1):17–26.