Simulating Spreading Activation: A Generative Approach to Semantic Similarity in Script-Variant Languages

Chia-Hsiang Ma[1], Yi-Ting Yu[1], Lee-Xieng Yang[1], Chia-Huei Tseng[2], Miao Cheng[2]
[1]Department of Psychology, National Chengchi University
[2]Research Institute of Electrical Communication, Tohoku University
{110752501, 112752502, lxyang}@nccu.edu.tw, {chia-huei.tseng.a8, cheng.miao.c3}@tohoku.ac.jp

Abstract

Multilingual embeddings often prioritize surface forms, struggling with script variants. We show "Identity Mismatch" in Japanese: models prefer English translations over *Kana* transcripts. We employ Generative Jaccard, using LLMs to simulate Spreading Activation. Though surface priming persists, it mitigates the identity gap and captures emotional nuances missed by static embeddings. We advocate shifting from static distance to dynamic generative overlap for better semantic modeling.

1 Introduction

Sentence similarity models, such as BERT and Sentence-BERT [1], are often assumed to capture deep semantics. However, recent studies suggest that they rely on "lexical overlap heuristics" [2] and are biased by word frequency [3]. This limitation is critical for Japanese, because the same concept can be represented in different ways (e.g., *Kanji* vs. *Kana*). If static embeddings merely encode surface patterns, the distinction between semantic cosine similarity and surface-level string metrics becomes blurred. This study investigates whether cosine similarity in BERT-based models is functionally dependent on the ratio of Longest Common Subsequence (LCS).

To address this issue, we employed Generative Jaccard [4], [5], based on the Spreading Activation Theory [6]. Instead of comparing static vectors, we utilize Large Language Models (LLMs) to generate token distributions to simulate cognitive retrieval. By treating the input as a "prime," this approach prioritizes intent over surface text.

We evaluate this method using 2,106 naturalistic scenario titles from the DIEM-A dataset [7]. Unlike synthetic adversarial examples, these texts were prepared by professional stage performers for bodily emotional expression. We demonstrate that while static Cosine similarity exhibits lexical bias (correlating with LCS), Generative Jaccard captures the semantic affinity of linguistically distinct but semantically identical pairs. This work bridges connectionist representations and cognitive theories, offering an alternative approach to measuring sentence similarity with LLMs.

2 Related Work

2.1. Vector Space Models and Cosine Similarity

The transition from discrete symbolic representations to distributed representations is central to modern NLP [8]. The common method to estimate semantic textual similarity (STS) involves encoding sentences into fixed-dimensional vectors using pre-trained language models, such as BERT, and computing the cosine similarity between them [1]. However, the "anisotropy" problem [9], reveals that contextualized representations occupy a narrow cone in the vector space rather than being uniformly distributed, which distorts similarity measures. Crucially, cosine similarity is systematically biased by word frequency [3]; it tends to underestimate the similarity of high-frequency words due to the anisotropic structure of the embedding space. Furthermore, raw BERT embeddings often fail to capture semantics without flow-based post-processing transformations [10]. These findings suggest that while cosine similarity is the standard, it may be influenced by statistical artifacts (such as frequency and length) rather than purely reflecting semantic closeness.

2.2. Surface-Level Metrics and Lexical Heuristics

Sentence similarity was primarily measured using

string distance metrics, such as the Jaccard similarity coefficient [11], which calculates the intersection over the union of token sets. While Jaccard is effective for tasks like name matching or identifying near-duplicates, it is limited by its inability to capture synonymy or paraphrase. Paradoxically, recent studies indicate that sophisticated neural models sometimes rely on similar surface-level strategies. Models like BERT often predict entailment based solely on high lexical overlap, which can be interpreted as Jaccard similarity, thus exhibiting "lexical overlap heuristics" [12]. This reliance on surface forms is particularly problematic in languages like Japanese, where different scripts (*Kanji*, *Hiragana*, *Katakana*) or phrasing can convey the exact same intent with no overlap at the character level. This raises the question of whether high cosine similarity scores in current similarity-based evaluations exactly reflect understanding, or merely capturing complex patterns of lexical co-occurrence.

2.3. Spreading Activation and Generative Expansion

To transcend the limitations of static vectors and surface matching, this study adopts the cognitive theory of Spreading Activation. This theory conceptualizes human semantic memory as a network in which activation of one concept spreads to related concepts [13]. In the context of LLMs, the text generation process can be viewed as a computational analog of spreading activation. This principle has been shown to be effective in Information Retrieval through document expansion [14]. A sequence-to-sequence model is used to predict potential queries for a document, enriching its representation with latent terms absent from the original text. This suggests that the "semantic neighborhood" of a sentence is better defined by the associations generated by the model, rather than by the static coordinates of the input sentence itself.

This research lies at the intersection of vector-based similarity, surface-level heuristics, and spreading activation. We hypothesize that Cosine similarity may inadvertently reflect lexical overlap rather than true semantic closeness. To address this, we employ Generative Jaccard, drawing on LLMs to expand the semantic representation of scenario titles. This enables similarity comparisons based on activated conceptual associations rather than static embeddings.

3 Methodology

3.1. Dataset: Large-Scale Naturalistic Scenarios

To evaluate the semantic sensitivity of embedding models in a realistic setting, we utilized the DIEM-A dataset [7]. For our study, we focused on the Japanese scenario titles, totaling 2,106 items. The dataset was originally developed for the collection of naturalistic bodily movement data. Professional stage performers designed the scenario titles and enacted corresponding movements in a controlled recording environment.

The data is balanced across 13 emotional categories (12 emotions such as Anger, Joy, and Grief, plus one Neutral condition), with each category containing 162 scenarios. To test the correlation between character overlap and similarity scores, we prepared four textual variations for each scenario: 1) Original Japanese (T_{JP}): The source text containing both *Kanji* and *Kana* characters. 2) *Kana* Transcript (T_{Kana}): The original Japanese text in *Kanji* converted to *Hiragana*, representing identical semantics and pronunciation but a distinct visual form. 3) Traditional Chinese (T_{ZH}): A translation into Traditional Chinese, preserving the logographic character system (*Hanzi*). 4) English (T_{EN}): A translation into English, using a distinct script system (Latin alphabet).

To ensure semantic equivalence across the 2,106 pairs per language, the translations underwent a structured validation process focusing on semantic correctness and linguistic fluency. The Traditional Chinese translations were independently cross-verified by two Chinese-Japanese evaluators with advanced proficiency in both languages (JLPT N1). The English translations were independently reviewed and refined by three evaluators with English proficiency at CEFR B2 level or above.

3.2. Static Embedding Models

To ensure a rigorous evaluation, we selected three widely used sentence embedding models for Semantic Textual Similarity (STS): Sonoisa [15], a Japanese-specific SBERT variant; E5-Multilingual [16], a contrastively trained multilingual embedding model; and MPNet [17], a strong multilingual baseline. These models represent competitive dense retrieval approaches beyond raw BERT baselines. For each model, we computed

sentence embeddings via mean pooling and calculated the cosine similarity between corresponding pairs.

3.3. Generative Jaccard

To operationalize the Spreading Activation theory [6] and decouple semantic measurement from surface constraints, we adopted the Generative Jaccard metric following previous works [4], [5]. We utilized Google's Gemma 3 (gemma-3-4b-it), an instruction-tuned LLM, to function as the cognitive activator.

This method treats the input sentence as a "prime" stimulus to trigger a conceptual network. We prompted the model to *"generate a list of 100 keywords, nouns, or related concepts"* associated with the input text. To assess cross-lingual semantic stability, we instructed the model to generate these keywords in three pivot languages: English, Traditional Chinese, and Japanese. To ensure determinism and reproducibility, the generation was performed using a greedy decoding strategy with a maximum of 512 new tokens.

Let T_A and T_B be the sets of unique token IDs generated by the model for sentences *A* and *B*, respectively. The similarity is defined as the Jaccard index of their generated conceptual fields:

$$Sim_{gen}(A, B) = \frac{|T_A \cap T_B|}{|T_A \cup T_B|}$$

This metric posits that if two sentences are semantically identical (e.g., share the same intent), they should trigger highly overlapping associative networks in the LLM's latent space, regardless of the script used in the input prompt.

3.4. Statistical Analysis

We adopted a two-stage statistical approach to examine the effects of script variation and surface overlap on similarity scores. First, we conducted a one-way repeated-measures ANOVA with script condition as an independent variable. For the Generative Jaccard metric, this analysis was extended to include three Pivot Languages (English, Traditional Chinese, Japanese) to assess the robustness of LLM-generated conceptual overlap across output languages. T_{JP}-T_{Kana} was treated as the identity baseline. Second, linear regression analyses were performed on the T_{JP}-T_{Kana} and T_{JP}-T_{ZH} pairs for both metric types, with similarity score as the dependent variable and Normalized LCS and Emotion Category as independent variables. Pearson's correlation coefficients were calculated to assess the linear relationship between character overlap and similarity scores.

4 Results

4.1. Lexical Bias in Static Cosine Similarity

We first examined the robustness of state-of-the-art static embeddings (E5-Multilingual and MPNet) against script variation. Since Sonoisa SBERT is a monolingual model, it was analyzed only in the regression phase. Cross-lingual comparisons (e.g., T_{JP} - T_{EN}) are outside its design scope.

The One-way Repeated Measures ANOVA revealed a significant main effect of script condition for both E5-Multilingual ($F(5, 10525) = 4856.7$, $p < .001$) and MPNet ($F(5, 10525) = 2668.7$, $p < .001$). This confirms that, despite careful translation quality control, the similarity judgments of these models vary significantly depending on the writing system used. In other words, they do not remain stable across semantically equivalent texts. Further analysis identified two distinct patterns driven by character overlap (LCS):

1) The "*Kanji* Bonus" (T_{JP}-T_{ZH} > T_{JP}-T_{EN}): Japanese-Chinese pairs received significantly higher similarity scores than Japanese-English pairs ($p < .001$) for both models. Regression analysis further indicated that this effect is predicted by character overlap: MPNet exhibited a significant positive association with normalized LCS ($\beta = 0.26$, $p < .001$), suggesting dependence on shared logographic characters (*Kanji*/*Hanzi*).

2) Identity Mismatch (T_{JP}-T_{Kana} < T_{JP}-T_{EN}): The T_{JP}-T_{Kana} condition can be considered linguistic identity. However, MPNet rated the English translation (T_{JP}-T_{EN}) as significantly more similar to the original text than the *Kana* transcription (T_{JP} - T_{Kana}) ($t = 31.9$, $p < .001$). Regression analysis on Sonoisa ($\beta = 0.86$, $p < .001$) and MPNet ($\beta = 0.73$, $p < .001$) revealed an extreme dependency on surface form, suggesting these models treat *Kana* transcripts as semantically distant simply due to the absence of visual *Kanji* overlap.

4.2. Spreading Activation in Generative Jaccard

We examined the Generative Jaccard metric using Gemma 3, which introduces an intermediate keyword generation stage in a specific pivot language. Separate one-way RM ANOVAs were performed for each pivot language (English, Chinese, Japanese) to test the stability of semantic retrieval.

When prompted in English ($F(5, 10525) = 166.0$, $p < .001$) or Traditional Chinese ($F(5, 10525) = 596.3$, $p < .001$), the results are consistent with the findings in static embeddings. Post-hoc comparisons revealed that T_{JP}-T_{ZH} pairs (sharing *Kanji*) had significantly higher similarity than T_{JP}-T_{EN} pairs (Pivot EN: $p < .001$; Pivot ZH: $p < .001$). This confirms that *Kanji* presence enhances retrieval even across pivot languages, which demonstrates the persistence of the Kanji Bonus (Pivot EN & ZH).

A distinct pattern emerged when using Japanese as the pivot language ($F(5, 10525) = 349.6$, $p < .001$). T_{JP}-T_{Kana} pairs showed strong resilience, with T_{JP} - T_{EN} scores higher than T_{JP} - T_{ZH} in some comparisons, and T_{JP} - T_{Kana} remained more connected to the original text than the Chinese pivot. Crucially, the identity gap observed in static models was mitigated, suggesting that matching output and input phonology (Japanese) makes spreading activation less inhibited by orthographic mismatch, allowing the model to bridge the *Kanji-Kana* gap. This pattern reflects the model's "context-sensitive" activation, showing how Generative Jaccard can bridge the *Kanji-Kana* gap and adapt its bias depending on activation paths.

The generative process is not immune to surface-level priming. Linear regression on T_{JP} - T_{Kana} pairs (via English pivot) revealed a significant positive correlation between LCS and Jaccard similarity ($\beta = 0.52$, $p < .001$). This confirms that the generated token distribution is still influenced by the input's surface form. Furthermore, the method captured subtle differences in Emotion Categories. In the pure semantic condition (T_{JP} - T_{EN}), Gemma 3 exhibited widespread sensitivity to emotional content. Alignment was significantly higher for discrete emotions (e.g., Anger, Joy, Shame) than for neutral texts ($p < .001$ for 11 out of 12 emotions). This suggests that emotional "primes" trigger more coherent and distinct associative networks across languages than neutral statements.

5 Discussion and Conclusion

Our findings reveal a critical "Surface Trap" in static sentence embeddings: models like MPNet often default to superficial string matching when encountering script variations, causing identity mismatch in Japanese retrieval. In contrast, the Generative Jaccard approach adopts a constructivist model of Spreading Activation. While surface priming persists (as evidenced by the *Kanji* Bonus), this method unfolds semantic nuances and shows that emotion acts as a "binding agent," creating denser conceptual clusters. Although computationally more intensive, this approach aligns more closely with human cognition. We advocate a paradigm shift: from measuring static distances between embeddings to evaluating dynamic overlaps in associative networks, to reconcile symbolic rigidity with fluid meaning in the LLM era.

Acknowledgements

We would like to thank Assistant Researcher Han-Wei Cheng from National Science and Technology Council for proofreading the Chinese translations, and graduate students Wan-Chen Chan and Yun-Ru Li, along with undergraduate student Zi-Rui Huang from Department of Psychology, National Chengchi University for their help in reviewing the English translations. Their input greatly improved the clarity of this work.

References

[1] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, November 2019.

[2] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3428–3448, July 2019.

[3] Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. Problems with Cosine as a Measure of

Embedding Similarity for High Frequency Words. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics** Vol. 2, pp. 401–423, May 2022.

[4] Chia-Hsiang Ma, Yi-Ting Yu, Lee-Xieng Yang, Chia-huei Tseng, and Miao Cheng. Culture Differences on Emotional Expression and Emotional Classification. In **Proceedings of the 13th International Conference on Affective Computing and Intelligent Interaction (ACII) Workshop on Multilingual and Multimodal Affective Computing (MMAC)**, Canberra, Australia, October 2025.

[5] Yi-Ting Yu, Chia-Hsiang Ma, Lee-Xieng Yang, Chia-huei Tseng, and Miao Cheng. Gender Differences on Textual Emotion Expression and Recognition. In **Proceedings of the 13th International Conference on Affective Computing and Intelligent Interaction (ACII) Workshop on Multilingual and Multimodal Affective Computing (MMAC)**, Canberra, Australia, October 2025.

[6] Allan M. Collins and Elizabeth F. Loftus. A Spreading-Activation Theory of Semantic Processing. **Psychological Review**, Vol. 82, No. 6, pp. 407–428, 1975.

[7] Cheng, M., Tseng, C. H., Fujiwara, K., Schneider, V., & Kitamura, Y. Asian Emotional Body Movement Database: Diverse Intercultural E-Motion Database of Asian Performers (DIEM-A). In **13th International Conference on Affective Computing and Intelligent Interaction**, Canberra, Australia, October, 2025.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. **Advances in Neural Information Processing Systems**, Vol. 26, pp. 3111–3119, December 2013.

[9] Kawin Ethayarajh. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 55–65, November 2019.

[10] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the Sentence Embeddings from Pre-trained Language Models. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 9119–9130, November 2020.

[11] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. In **Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)**, pp. 73–78, August 2003.

[12] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3428–3448, July 2019.

[13] Allan M. Collins and Elizabeth F. Loftus. A Spreading-Activation Theory of Semantic Processing. **Psychological Review**, Vol. 82, No. 6, pp. 407–428, 1975.

[14] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document Expansion by Query Prediction. **arXiv preprint arXiv:1904.08375**, September 2019.

[15] Sonoisa. sentence-bert-base-ja-mean-tokens-v2. Hugging Face, 2021. [Pre-trained model]. Available: https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

[16] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 Text Embeddings: A Technical Report. **arXiv preprint arXiv:2402.05672**, February 2024.

[17] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and Permuted Pre-training for Language Understanding. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 16875–16886, December 2020.