

RAG 時代の言語資源設計原理 — 構造化テキストによる「参照可能性」の実証

池松 潤

概要

本研究は、RAG 時代における言語資源の新しい設計原理を提案し実証する。従来の言語資源は「分類精度」で評価されてきたが、RAG の普及により AI が参照先として到達できるかという「参照可能性」が重要な評価軸となる。本研究では、連載中エッセイ「嫉妬マニア」を原典とする嫉妬 AI 辞書 (321 件) を構築し、構造化メタデータのみでの表現と本文のみの表現の検索性能を比較した。実験の結果、構造クエリにおいて構造化メタデータのみ (59.0%) は本文のみ (5.3%) に対して Recall@10 で 11.1 倍の性能差を示した ($p < 0.001$)。これは言語資源を「人間が読む文章」から「AI が参照する仕様書」へ設計転換することの有効性を実証する。

1 はじめに

1.1 本研究の立場

本研究は「感情を扱う応用研究」ではない。本研究の目的は、大規模言語モデル (LLM) と RAG が普及した現代において、AI が参照せざるを得ない言語資源の新しい設計原理を確立することにある。「嫉妬」は実証の題材であり、本質は言語資源の設計原理にある。

RAG の普及により、言語資源の価値基準は変化しつつある。従来は分類器の訓練データとしての「分類精度への貢献」が主要な評価軸であったが、RAG においては「クエリに対して正しい参照点へ到達できるか」という**参照可能性 (Referenceability)** が本質的な価値となる。

本研究で構築したデータセットは、原典テキストを「読み物」から「仕様書」へ変換したものである。すなわち、人間が通読するための文章ではなく、AI が参照構造として利用するために設計された資源である。



図1 構造化メタデータによる参照可能性の向上

1.2 研究の問い

本研究は以下の問いに答える：構造化メタデータを中心とした文書表現は、本文 (自由文) を含む表現より参照可能性が高いか。

1.3 本論文の構成

2 節で関連研究、3 節でデータセット設計、4 節で実験設定、5 節で結果、6 節で考察、7 節で結論を述べる。

本論文を読む際、以下の誤読を避けていただきたい：(1)「感情分類データセットの研究」ではない。嫉妬は実証の題材であり、本質は言語資源の設計原理にある。(2)「欠損だらけのデータ」ではない。空欄は設計上の意図的な判断である (3.3 節参照)。

2 関連研究

RAG[1] は外部知識を検索して生成に活用する手法であり、REPLUG[2] やノイズの影響研究 [3] 等で発展している。これらの研究はアルゴリズムの改善に焦点を当てるが、本研究は**データ設計**による性能向上を目指す点で異なる。

感情に関する言語資源として、英語では NRC Emotion Lexicon[4]、日本語では評価表現辞書 [5] や感情表現辞典 [6] がある。しかし、これらは「嫉妬」のような複合感情を扱わない。Google の GoEmotions は 27 種類の感情カテゴリを定義するが、「嫉妬 (jealousy)」は含まれていない。複合感情の取扱いが困難であるからと思われる。

近年，Song et al.[7] は Large Emotional World Model (LEWM) を提案し，感情情報を動的にモデル化することで推論性能が向上することを示した．本研究の嫉妬 AI 辞書は静的な参照点として機能し，LEWM の動的モデルに対して文化固有の感情構造を提供できる点で相補的である．

3 データセット設計

本論文では，作品本文から抽出・構造化した要約テキストや感情タグ等を総称して「メタデータ」とする

3.1 データセット概要

データセット名は嫉妬 AI 辞書 (Jealousy AI Dictionary)．原典は婦人公論.jp 連載中「嫉妬マニア」(著者：斉藤ナミ氏，2025 年 3 月～12 月)である．連載中の記事を研究サンプルとして使用する点は極めて珍しい．総エピソード数 321 件，評価用クエリ 400 件．アノテーションは筆者が実施し，斉藤氏に全エピソードのラベル確認およびメタデータのサンプル確認をいただいた．

3.2 構造化メタデータ (25 項目)

各エピソードに 25 項目のメタデータを付与した．実験では，検索に有効な 9 項目を連結した core_text を使用した：嫉妬の方向，分析対象，嫉妬対象，嫉妬語彙，嫉妬構造カテゴリ，比較の構図，心理的距離，努力介入可能性，行動パターン．これらは統制された語彙で記述され，検索クエリとの語彙マッチングを向上させる．

特に primary_emotion_vector の設計は重要である．GoEmotions に「嫉妬」がない理由は，複合感情の取り扱いが難しいためである．本項目では，複数感情の強度を 0.0～1.0 の連続値で JSON 形式表現する：

```
{"anger": 0.8, "sadness": 0.6, "desire_for_recognition": 0.9}
```

本形式は感情の混合状態 (感情のレシピ) を AI に学習させることを可能にする．英語では「嫉妬」「羨望」「妬み」の区別が曖昧だが，日本語では明確に異なる．このような日本語特有の感情構造を構造化することは，ソブリン AI 開発への貢献となる．

3.3 空欄の設計 (戦略的スキップ)

本データセットの空欄は欠損ではなく，戦略的・意図的な空欄である．実験の結果，空欄を「なし」等の文字列で埋めるより，空欄項目をスキップする方が高い検索性能が出る．不確実な情報を無理に埋め込むよりも，確実な情報のみを出す方が RAG の参照精度が高まるためである．

4 実験設定

4.1 タスクと評価指標

タスクは Reference Retrieval (参照点回収)：クエリに対し正解エピソードを上位 k 件に返す．評価指標は Recall@10 (R@10)．検索モデルは TF-IDF + コサイン類似度 (char 3-5gram)．

4.2 文書表現バリエーション

4 条件を比較 (表 1)：

条件名	内容
core_text	構造化メタデータのみ (9 要素を連結)
long_text_only	本文のみ (構造化なし)
core_plus_long	構造化メタデータ + 本文
short_text	短い要約のみ

4.3 クエリ設計

400 件のクエリを生成：shorttext (100 件，原文抜粋)，comparison (100 件，比較構図)，distance_effort (100 件，心理的距離・努力介入)，behavior (100 件，行動パターン)．

評価分割として，STRUCT (構造クエリ 300 件：comparison, distance_effort, behavior) と SHORT (原文抜粋 100 件) を設定．STRUCT クエリは構造化メタデータの語彙を用いた検索，SHORT クエリは原文の語彙を用いた検索である．

5 結果



図2 構造クエリ (STRUCT, $n=300$) における検索性能

5.1 主結果

図2および表2に主結果を示す。構造クエリ (STRUCT, $n=300$) において, core_text (構造化メタデータのみ) は $R@10=59.0\%$ を達成し, long_text_only (本文のみ, 5.3%) に対して **53.7ポイント**, **11.1倍**の改善を示した。

表2 クエリスタイル別の Recall@10

クエリスタイル	core_text	long_text	差	倍率
comparison	61%	5%	+56pt	12.2倍
distance_effort	51%	7%	+44pt	7.3倍
behavior	65%	4%	+61pt	16.3倍
STRUCT 平均	59%	5.3%	+53.7pt	11.1倍
shorttext	100%	100%	0pt	1.0倍

5.2 統計的有意性

McNemar 検定の結果, core_text vs long_text_only の差は統計的に極めて有意であった ($p < 0.001$)。

5.3 本文追加の効果

図2のとおり, core_text に long_text を追加した core_plus_long は 58.3% であり, core_text (59.0%) と同等であった。すなわち, **本文を追加しても精度は向上しない**。

6 考察

6.1 構造化メタデータの優位性

構造化メタデータのみでの表現が 11.1 倍優れた理由: (1) ノイズ排除: 本文には検索クエリと直接対

応しない表現が含まれ TF-IDF でノイズとなる, (2) 語彙一貫性: 統制された語彙によりクエリとのマッチング向上, (3) 空欄スキップ: 有意味情報の密度向上。

6.2 「量は質ではない。純度こそ質である」

本結果は, 言語資源設計において「量より純度」が重要であることを示す。321 件という小規模データセットでも, 構造化による純度向上で 11 倍の性能差を実現した。

6.3 他ドメインへの一般化

「嫉妬」は特殊な感情に見えるが, 本設計原理は他ドメインへ一般化可能である。例えば, ハラスメント防止による離職者対策 (HR 領域), メンタルヘルスケアのカスタマーサポート補助, 経営者 (CEO) デジタルツインによる事業哲学の次世代継承などへの応用が期待できる。

6.4 限界

本研究は TF-IDF + コサイン類似度を採用した。これは語彙マッチングの効果を直接観察し, 構造化メタデータによる語彙統制の有効性を検証するための意図的な選択である。埋め込みベース手法では語彙の表層一致に依存しないため, 構造化の効果が異なる形で現れる可能性がある。Dense retriever での検証, および嫉妬以外のドメインへの実証的検証は今後の課題である。

7 結論

本研究は, RAG 時代の言語資源設計において「参照可能性」という新しい評価軸を提案し, 構造化メタデータ中心の設計が有効であることを実証した。

主要知見: (1) 構造化メタデータのみでの表現は本文のみでの表現より 11.1 倍優れた ($59\% vs 5.3\%$, $p < 0.001$), (2) 本文追加で性能は向上しない, (3) 空欄スキップ方式は有効。

言語資源を「読み物」から「仕様書」へ設計転換する本原理は, 他ドメインへの一般化が期待される。データセットは 2026 年 3 月に [GitHub^{1\)}](https://github.com/junikematsu/shitto-mania-dic) および [Hugging Face^{2\)}](https://huggingface.co/datasets/samuraijun/shitto-mania-dic) で公開予定である。

1) <https://github.com/junikematsu/shitto-mania-dic>

2) <https://huggingface.co/datasets/samuraijun/shitto-mania-dic>

謝辞

齊藤ナミ氏には、連載『嫉妬マニア』の研究利用許諾、全エピソードのラベル確認、メタデータのサンプル確認を通じ、アノテーション品質確認など多大なるご協力をいただいた。

株式会社中央公論新社、ならびに婦人公論.jp 川口由貴編集長には、ご理解ご協力を賜った。記して感謝したい。

参考文献

- [1] Patrick Lewis, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.
- [2] Weijia Shi, et al. REPLUG: Retrieval-augmented black-box language models. arXiv:2301.12652, 2023.
- [3] Florin Cuconasu, et al. The power of noise: Redefining retrieval for RAG systems. arXiv:2401.14887, 2024.
- [4] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), 2013.
- [5] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. *自然言語処理*, 12(3), 2005.
- [6] 中村明. 感情表現辞典. 東京堂出版, 2004.
- [7] Changhao Song, et al. Large Emotional World Model. arXiv:2512.24149, 2025.

付録：実験に使用したメタデータ項目

表3にデータセット25項目と、実験（core_text条件）で使用した9項目を示す。

これら9項目を連結し、空欄項目をスキップしたものがcore_text条件である。25項目と実験で利用した9項目を一覧にする。

表3 データセット項目一覧（25項目）とcore_text使用9項目（ ）

#	項目名（英）	説明（日）	core_text
1	id	番号	
2	date	掲載日	
3	source_text	記事タイトル	
4	jealousy_direction	嫉妬の方向	
5	analysis_target	分析対象	
6	target_of_jealousy	嫉妬の対象	
7	jealousy_vocabulary	嫉妬の語彙・名言	
8	jealousy_structure_category	嫉妬の構造	
9	context_tag	文脈タグ	
10	comparison_pair	比較二項対	
11	analysis_axis_distance	分析軸：距離	
12	analysis_axis_effort_intervention	分析軸：努力介入	
13	hierarchy_source	階層（発生源）	
14	behavior_pattern	行動類型	
15	primary_emotion_vector	主要感情ベクトル	
16	consciousness_state_tag	意識状態タグ	
17	embodied_cognition_text	身体化認知テキスト	
18	entry_id	エピソードID	
19	column_id	元コラムID	
20	paragraph_span	段落・文範囲	
21	short_text	短い要約文	
22	long_text	原文抜粋	
23	narrative_role	語り手の立場	
24	span_highlight	嫉妬ハイライト	
25	source_url	記事のURL	