「〇〇で歌ってみた」替え歌を用いた音韻類似単語検索ベンチマークの構築

島谷 二郎 (個人)

2025/03/14 (Fri)
言語処理学会第31回年次大会 併設ワークショップ JLR2025
「日本語言語資源の構築と利用性の向上」

データセットURL

「〇〇で歌ってみた」替え歌に基づく音韻検索データセットを構築

- 貢献
  - 言語ユーモアに基づく日本語音韻検索データセットの作成と公開

- 本発表では
  - 背景：なぜ作ったか
  - 構築：どう作ったか
  - 検証：どう使えそうか

について報告


```
{
 "queries": [
  {"query": "アウマル","positive": ["タクマル"]},
  {"query": "アケ", "positive": ["アベ", "カケイ"]},
  ...
 ],
 "words": [
  "アーノルド",
  "クリスアーノルド",
  "アーリン",
  "ロビーアーリン",
  "アイ",
  "アイエイジロウ",
  ...
 ]
}
```


クエリ150件
- クエリが元歌詞
- 正例が替え歌

候補語13076件

目次

- **背景**
- 構築
- 検証
- まとめ

背景

公開データセットによって音韻検索の手法を比較可能にし分野全体での知見を蓄積しやすくしたい

- 音韻検索とは
- 音韻検索の評価の課題
- 音韻検索データセットの公開状況
- 目指す姿
- 既存研究: 駄洒落データベース
- 「〇〇で歌ってみた」

音韻検索とは

テキストの音韻の近さに基づく検索

- 検索例（クエリ：洗濯機）
  - 意味検索 -> ランドリー、乾燥機、ドラム
  - 音韻検索 -> 選択肢、ゲン担ぎ、ケンタッキー

- 応用例
  - 言語ユーモア: 駄洒落[1]、ラップ[2]、空耳[3]、言い間違いボケ[4]など
  - 音声対話: 音声認識誤りに頑健な検索[5]、Entityリンキング[6]など

1. 荒木健治. "駄洒落データベースを用いた駄洒落生成システムの性能評価." 人工知能学会第 2 種研究会: ことば工学研究会資料, SIG-LSE-B 703 (2018): 8.
2. 三林亮太, et al. "ラップバトルにおける逆向き生成によるライムを含む返答バース生成." *情報処理学会論文誌データベース (TOD)* 17.2 (2024): 28-39.
3. Shimaya, Jiro, Nao Hanyu, and Yutaka Nakamura. "Automatic generation of homophonic transformation for Japanese wordplay based on edit distance and phrase breaks." HCI International 2019-Posters: 21st International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21. Springer International Publishing, 2019.
4. 高山宜之, 原口 和貴, 北村達也, and 灘本明代, “音素類似関係を用いた言い間違えボケの自動生成”, 第13回データ工学と情報マネジメントに関するフォーラム(2021)
5. Zhou, Xiaozhou, Ruying Bao, and William M. Campbell. "Phonetic embedding for ASR robustness in entity resolution." (2022).
6. 邊土名朝飛, et al. "ユーザ発話とEntityの音声類似度を考慮した Entity Linking 手法の検討." 人工知能学会全国大会論文集 第 36 回 (2022). 一般社団法人 人工知能学会, 2022.

音韻検索の評価の課題

手法と評価方法がバラバラで横断的な知見が得られにくい

分野	研究	音韻検索手法	評価方法
言語ユーモア生成	駄洒落[荒木2018]	表層の一致から段階的に条件を緩めて検索	生成品質（検索精度は未評価）
	空耳[Shimaya2019]	母音と子音の編集距離	
	言い間違えボケ[高山2021]	音素特徴に基づく類似度	
	ラップ[三林2024]	母音の一致	
音声対話システム	検索[Zhou 2022]	音韻埋め込み	独自データセット（非公開）
	Entityリンキング [邊土名2022]	音素列の編集距離とセミグローバルアライメント	

＊引用研究は前ページと同様

音韻検索データセットの公開状況

音韻検索評価に活用しやすい公開データセットが不足している
駄洒落データベースも検索用ではない

分野	研究	データセット	公開状況
言語ユーモア生成	駄洒落[荒木2018]	駄洒落データベース[1]	公開
	空耳[Shimaya2019]	(なし)	-
	言い間違えボケ[高山2021]	(なし)	-
	ラップ[三林2024, Kawahara2007[1]]	ラップの書き起こし	非公開
音声対話システム	検索[Zhou2022]	動画と書籍の音声検索ログ（英語）	非公開
	Entityリンキング [邊土名2022]	音声対話アプリのEntity辞書とログ	非公開

1. 荒木健治, et al. "駄洒落データベースの拡張及び分析." *人工知能学会第 2 種研究会ことば工学研究会資料* (2018): 1-15.
2. Kawahara, Shigeto. "Half rhymes in Japanese rap lyrics and knowledge of similarity." *Journal of East Asian Linguistics* 16 (2007): 113-144.

上記以外の引用研究は前前ページと同様

目指す姿

公開データセットによって音韻検索の手法を比較可能にし分野全体での知見を蓄積しやすくする

現状

- 研究ごとに独自の手法を独自のデータセットやタスクで評価
- データセットは非公開のことも多い

→分野全体での知見が蓄積されにくい

理想

- 音韻検索の公開データセットで手法が比較可能
  （c.f. 意味検索分野では一般的）

手法	データセット1	データセット2	…
埋め込み1	0.78	0.33	
埋め込み2	0.83	0.30	
…	…	…	
編集距離1	0.65	0.55	
編集距離2	0.70	0.58	
…	…	…	

関連研究: 駄洒落データベース[1]

音韻検索における正例の集合として使えるが 文脈依存性や候補単語の不明確さが課題

- 概要
  - 駄洒落文67000件
  - 駄洒落の種表現、変形表現、種別、スコア評価が付与

・フォーマット
通し番号,原形,タグ付,種別,スコア１,スコア２,スコア３,平均スコア

・例
1,坊っちゃんがぼっちゃんと水に飛び込む,(坊っちゃん) が [ぼっ ちゃんと] 水 に 飛び込む,1,4,2,2,2.67
2,高菜、あったかな？,(高菜) 、 [あっ た か な] ？ ,1,3,3,3,3.00
3,炭のすみか,(炭) の [すみか],1,2,2,2,2.00
4,この寺の檀家はダンカン,この 寺 の (檀家) は [ダンカン],1,2,4,3,3.00
5,ナウシカを誘う鹿,(ナウ シカ) を [誘う 鹿],2,2,4,1,2.33
6,りんご園では、燐<りん>５円,(りんご 園) で は 、 [燐 <りん> 5 円],1,3,3,2,2.67

http://arakilab.media.eng.hokudai.ac.jp/~araki/dajare.htm

- 音韻検索データセット化するうえでの課題
  - 文脈依存性: 音韻だけでなく文章の自然さも単語の選択に影響
  - 候補語が不明確: 検索タスクにしにくい

1. 荒木健治, et al. "駄洒落データベースの拡張及び分析." *人工知能学会第 2 種研究会ことば工学研究会資料* (2018): 1-15.

「〇〇で歌ってみた」

特定ジャンルの単語で元歌詞の音韻を再現する替え歌

- 例（野球選手）
  - 元歌詞: うさぎ おいし かの やま
  - 替え歌: 宇佐美　大石　　鹿野 八名
- 特徴
  - 文章の自然さよりも音韻を重視
  - 候補単語が限定的

  → 音韻検索データセットの構築に有用

	駄洒落	〇〇で歌ってみた
文章の自然さ	必要	不要
候補単語	不明確	明確

YouTubeに投稿された動画の例

野球選手[1]

駅名[2]

国名[3]

魚[4]

サッカー選手[5]

ポケモン[6]

YouTuber[7]

四股名[8]

将棋棋士[9]

1. 「前前前世」を野球選手名で歌ってみた【君の名は。】 https://www.youtube.com/watch?v=H_TamoTCUJo
2. 【駅名替え歌】【全県6周】駅名で「背景ドッペルゲンガー」【Vo.重音テト】 https://www.youtube.com/watch?v=kxaHpG43WQ0
3. 【替え歌】国名だけで「lemon」歌ってみた【米津玄師】【ゆっくりが歌う】 https://www.youtube.com/watch?v=DN6504USVaE
4. 魚の名前で紅蓮華【鬼滅の刃】 https://www.youtube.com/watch?v=SmdG67C4XOk
5. 「メルト」をサッカー選手の名前だけで歌ってみた https://www.youtube.com/watch?v=-FVB6ZNR95Q
6. 【パプリカ】剣盾・初登場ポケモンだけでパプリカ【米津玄師×Foorin】 https://www.youtube.com/watch?app=desktop&v=JnpvoAk3XT4&t=30s
7. 【棋士の名は。】将棋棋士の名前で前前前世【垣瀬内シリーズ】 https://www.youtube.com/watch?v=vKX6WwuNAIE
8. 【替え歌】『YouTubeテーマソング』をYouTuberの名前だけで歌ってみた https://www.youtube.com/watch?v=nBcznBtiqA0
9. 砂の惑星 力士の四股名だけで歌ってみた 【ワトイモロー反射】 https://www.youtube.com/watch?v=SC1Ln57fA2o

目次

- 背景
- **構築**
- 検証
- まとめ

構築

「〇〇で歌ってみた」を解析し音韻検索データセットを構築

- 元データ
- 文章解析
- 単語アライメント
- フィルタリング
- 成果

元データ

効率化のためコンクールの提出作品（テキスト化済み）を活用

- 替え歌
  - 「野球選手名で歌ってみた」
  - 2023年開催の有志コンクールに提出されたテキストファイル46作品
    - 替え歌単語数約4400
    - 作詞者数は13（トーナメント制のため）
    - 曲の種類数は6（課題曲制のため）
- 候補単語リスト
  - 2024年時点での日本プロ野球在籍選手（OB含む）の姓、登録名、フルネーム

替え歌テキストサンプル

宇佐美 小石博孝
うさぎ追いしかの山

小塚 辻勇夫 河
こぶな釣りしかの川

シューメーカー 熊野輝光
夢は 今もめぐりて

夏目隆司 古田 荘
わすれがたきふるさと

文章解析→単語アライメント→フィルタリング

元歌詞の音韻カナと文節位置、替え歌歌詞の音韻カナを推定

生データ

宇佐美 小石博孝
うさぎ追いしかの山

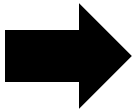
小塚 辻勇夫 河
こぶな釣りしかの川

シューメーカー 熊野輝光
夢は 今もめぐりて

夏目隆司 古田 荘
わすれがたきふるさと

音韻カナと文節の推定

形態素解析(SudachiPyのA単位）で
推定後目視で修正



音韻カナデータ（文節つき）

ウサミ コイシヒロタカ
ウサギ/p オイ シ カノ/p ヤマ/p

コヅカ ツジイサオ カワ
コブナ/p ツリ/p シ カノ/p カワ/p

シューメーカー クマノテルミツ
ユメ/p ワ イマ/p モ メグリ/p テ

ナツメタカシ フルタ ソウ
ワスレ/p ガタ キ フルサト/p

＊「/p」は文節開始

替え歌単語と対応する元歌詞の分割を推定

音韻カナデータ

ウサミ コイシヒロタカ
ウサギ オイ シ カノ ヤマ

コヅカ ツジイサオ カワ
コブナ ツリ シ カノ カワ

シューメーカー クマノテルミツ
ユメ ワ イマ モ メグリ テ

ナツメタカシ フルタ ソウ
ワスレ ガタ キ フルサト

替え歌単語を基準とした系列アライメント

- 各ペアの類似度スコアを最大化する元歌詞の分割位置を動的計画法で計算
- 類似度スコアは母音と子音の一致率をベースに結果を見ながら複数回調整

単語アライメントデータ

[“ウサミ”, “コイシヒロタカ”]
[“ウサギ”, “オイシカノヤマ”]

[“コヅカ”, “ツジイサオ”, “カワ”]
[“コブナ”, “ツリシカノ”, “カワ”]

[“シューメーカー”, “クマノテルミツ”]
[“ユメワ”, “イマモメグリテ”]

[“ナツメタカシ”, “フルタ”, “ソウ”]
[“ワスレガタキ”, “フルサ”, “ト”]

文章解析→単語アライメント→**フィルタリング**

頻度が多く、元歌詞の分割が不自然でないペアを抽出

単語アライメントデータ
(ペアの頻度を集計)

元歌詞	替え歌歌詞	頻度	文節開始	単語終了
ウサギ	ウサミ	3	true	true
オイシカノヤマ	コイシヒロタカ	1	true	true
コブナ	コヅカ	3	true	true
ツリシカノ	ツジイサオ	2	true	true
…	…	…	…	…

以下の条件でフィルタ

- 頻度≧3
- 文節開始=true
- 単語終了=true

クエリ-正解データ

```
[
  {“query”: “ウサギ”, “positive”: [“ウサミ”]},
  {“query”: “コブナ”, “positive”: [“コヅカ”]},
  …
]
```

成果

替え歌46作品（約4400件の元歌詞-替え歌単語ペア）からクエリ150件を抽出
候補単語リスト13076件と合わせてデータセット化

クエリ150件
- クエリが元歌詞
- 正例が替え歌

候補単語リスト13076件
- 日本プロ野球（NPB）在籍選手の姓、登録名、フルネームのカナ表記

```
{
 "queries": [
  {"query": "アイ", "positive": ["アイ"]},
  {"query": "アウマル","positive": ["タクマル"]},
  {"query": "アケ", "positive": ["アベ", "カケイ"]},
  ...
 ],
 "words": [
  "アーノルド",
  "クリスアーノルド",
  "アーリン",
  "ロビーアーリン",
  "アイ",
  "アイエイジロウ",
  ...
 ]
}
```

目次

- 背景
- 構築
- **検証**
- まとめ

検証

作成したデータセットを妥当性や傾向を評価

- 精度が想定できる音韻類似度指標を用いた妥当性検証
- 目視（検索結果との比較）
- 目視（音声認識誤りとの比較）
- 母音の重要度
- LLMによる音韻検索

検証

作成したデータセットを妥当性や傾向を評価

- 精度が想定できる音韻類似度指標を用いた妥当性検証
- 目視（検索結果との比較）
- 目視（音声認識誤りとの比較）
- 母音の重要度
- LLMによる音韻検索

精度が想定できる音韻類似度指標を用いた妥当性検証

想定通りの結果であったためデータセットの内容はある程度妥当と判断

音韻類似度	**説明**	**精度（想定）**	**精度（結果）**	
			recall@1	recall@10
カナ	カナの編集距離	低い	0.168	0.455
音素	音素の編集距離	やや高い	0.341	0.672
子音＋母音	子音と母音それぞれの編集距離の和	やや高い	0.381	0.676
音響モデル	子音と母音それぞれの音響モデルに基づく重み付き編集距離の和	高い	0.420	0.788

目視（検索結果との比較）

概ね妥当だが、素人目には正例と甲乙つけがたい検索結果もあった
元音源の歌い方や作詞者の嗜好が影響している可能性

クエリ	**上位検索結果**（太字は正例、下線は甲乙つけがたい検索結果）
ソバデ	ソガベ, **ソラーテ**, ソノベ, トガメ, トバ,
ドレ	トベ, **トーベ**, トネ, ウォーレン, ジョウベ
ナイタ	マイタ, ダイタ, **ナリタ**, ナイト, ナニータ
ナニガ	ナニータ, **ナリタ**, ナシダ, アリガ, タニダ
デモ	レモン, **デイモン**, デード, セノ, クモン
モドラ	モトダ, ラドラ, **ボトラ**, モトハラ, ハドラー

考えられる対策
- 替え歌作詞者によるチェック
- データ数を増やす

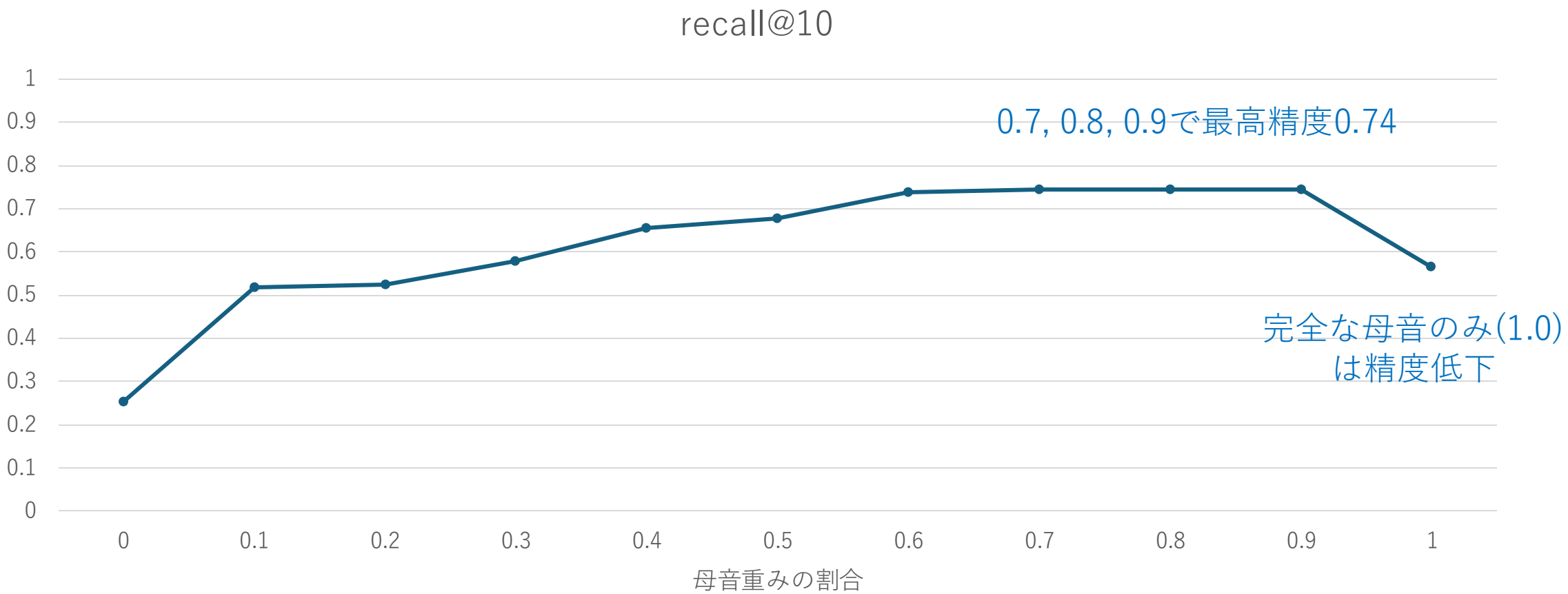
目視（音声認識誤りの音韻ペアとの比較）

音声認識誤りと「〇〇で歌ってみた」で音韻ペアの傾向が多少異なる
本データセットの汎用性検証は今後の課題

「〇〇で歌ってみた」（歌唱）	音声認識誤り（発話）
モウラ数を保持しつつ音を置換 例 • 「タメニ→カメニ」（×「タニ」） • 「ママ→ハマ」（×「マエマ」） 歌のリズム構造と韻を重視	モウラの削除が頻繁 例 • 「下剋上→下女」 • 「殺風景→風景」 音声波形の類似性を重視

母音の重要度

母音重視（~0.8）で最高精度、完全な母音のみは精度低下



LLMによる音韻検索

LLMに音韻検索精度は編集距離（子音＋母音）より低い
汎用的なLLMの音韻理解能力がまだ低いことを示唆

- 方法
  - 編集距離（子音＋母音）のtop100と正例をLLMに与え
    上位10件を推測させた
- 結果
  - 編集距離（子音＋母音）より
    低い精度
  - LLMの中ではgpt-4.5が高い

条件	recall@10
編集距離（子音＋母音）	0.676
gpt-4o-mini	0.444
gpt-4o	0.508
gemini-2.0-flash	0.496
gpt-4.5-preview	0.583

参考: システムプロンプト

音韻類似度の説明と例示を含むシステムプロンプトを使用
出力は構造化モードでindexのリストを取得

音韻類似度の説明

```
You are a phonetic search assistant.
You are given a query and a list of words.
You need to rerank the words based on phonetic similarity to the query.
When estimating phonetic similarity, please consider the following:
1. Prioritize matching vowels
2. Substitution, insertion, or deletion of nasal sounds, geminate consonants, and long vowels is acceptable
3. For other cases, words with similar mora counts are preferred
You need to return only the reranked list of index numbers of the words, no other text.
You need to return only topn index numbers.
```

例示

```
Example:
Query: タロウ
Wordlist:
0. アオ
1. アオウゾ
2. アノウ
3. タキョウ
4. タド
5. タノ
6. タロウ
7. タンノ
Top N: 5
Reranked: 6, 4, 5, 7, 2
```

参考: LLMの音韻検索失敗事例

文字の表層的な一致を重視してしまい
母音やモウラ数の一致を優先したリランクができていない

- 例1

クエリ	オサキモ
検索結果	オザキ, オサカ, オサダ, オサナイ, オサリ, オダキミオ, オカジマ, オダジマ, オジマ, オゼキ
正例	ロサリオ, トマシノ

- 例2

クエリ	モグリ
検索結果	モギ, モリ, コモリ, モリグチ, モトキ, モリキ, ノグチ, ノガミ, モテギ, モモイ
正例	モスビー

参考: データセット公開

構築したデータセットをGitHubで公開（ライセンス: CDLA Permissive 2.0）

- https://github.com/jiroshimaya/soramimi-phonetic-search-dataset
- pipでインストールし、独自関数でrecallを計算可能

インストール

```
pip install soramimi-phonetic-search-dataset
```

基本的な使用例

```
from soramimi_phonetic_search_dataset import evaluate_ranking_function

# カスタムのランキング関数を定義
def my_ranking_function(query_texts: list[str], wordlist_texts: list[str]) -> list[list[str]
    # ここにあなたの音韻的類似度に基づくランキングロジックを実装
    return ranked_wordlists

# 評価の実行
recall = evaluate_ranking_function(ranking_func=my_ranking_function, topn=10)
print(f"Recall@10: {recall}")
```

まとめ

「〇〇で歌ってみた」替え歌に基づく音韻検索データセットを作成

- 成果
  - クエリ150件・候補語13076件の音韻検索データセット構築
  - 定量評価と目視確認で実用性を検証

  →音韻検索の手法比較や知見の蓄積が活発になることを期待

- 今後の課題
  - データ量増加: 動画を効率的に解析するパイプライン開発
  - 汎用性検証: 他の言語ユーモアや対話システム向けの音韻検索の評価に使えるか検証