

二種類の日本語読み推定タスク・データセット の構築

白井清昭

北陸先端科学技術大学院大学

背景

- 日本語の単語には読みの曖昧性がある
- 読み推定タスク
 - 特定の文に出現した漢字表記の単語の読みを推定する
 - 読みの候補はあらかじめ定義されている
 - ex. その頃彼女はすでに30歳を越えていたが、ローザノフは一目見るなり恋に落ちた。
{ ヒトメ, イチモク } → ヒトメ
- テキスト読み上げシステムなどで必要とされる要素技術

目的

- 日本語読み推定タスクのデータセットの構築・公開
 - 人手で正解の読みが付与された例文の集合
 - 読み推定モデルの学習や評価に利用できる
 - 公開されているものはない
 - 本プロジェクトの目的(のひとつ)
- 二種類のデータセット
 - 標本抽出データセット
 - 読みの分布は実際の使用分布に近い (しばしば偏りがある)
 - 標本均等化データセット
 - 読みの分布は均等に近い

関連研究

- 佐藤+ 2022
 - 読みの注釈付きデータから大規模な仮名注釈付きコーパスを構築
 - 国会図書館の書誌データ(本のタイトル) 3.4億文字
 - 青空文庫+点字データ「サビエ」 5453万文字
 - 読みのクラス分類の実験
 - BERTを使って系列ラベリング問題を解く
 - 正解率は0.77~0.99
 - 大規模
 - 読みは人手で付与されたもの
 - 大部分は本のタイトルで自然な文ではない

概要

- 目標とする日本語読み推定タスクのデータセット
 - 対象単語: 50個
 - 読みの曖昧性のある単語
 - 用例: 1単語当たり100件(以上)
 - 対象単語を含む文をテキストから抽出
 - アノテーション: 正解の読み
 - 2種類のデータセット
 - 標本抽出データセット ⊂ 標本均等化データセット
(読みの分布は実際の使用分布) (読みの分布は均等に近い)
 - モデルの学習ではなくシステムの評価向け

構築手順

1. 対象単語の選定
2. 用例の抽出
3. 読みの付与
 - 標本抽出データセットの構築
4. 低頻度の読みの用例の追加
 - 標本均衡化データセットの構築

昨年度のWSで発表

対象単語の選定

- 先行研究から候補語を選定
 - (小林・古宮 2021) 68語 (佐藤+ 2022) 205語
- 選定基準
 - 基本的に文脈依存の読みの曖昧性があるものを選択
 - ほぼ同じ意味を持ち、互いに置き換え可能なものは選ばない (異なる読みでも実体は同じ)
 - ×悪阻: オソ, ツワリ / ×怒る: オコル, イカル
 - 固有名詞の読みは選ばない (ex. 長尾: ○チョウビ, ×ナガオ)
 - 複合名詞の一部になる読みは選ばない (ex. 下手人: ×ゲシュ)
- 50個の対象単語と読みを選定
 - 一目(ヒトメ, イチモク)、上手(ウワテ, ジョウズ) など

用例の抽出

- 対象テキスト

- 日本語Wikipedia (2023-05-01版ダンプデータ)
- 大規模、分野が多様、公開時に引用元が示しやすい

- 前処理

- 本文の抽出
- 文分割(「。」 「.」 「？」 「！」 で分割)
- 形態素解析 MeCab + UniDic

- 用例の抽出

- 書字形基本形が一致する単語とそれを含む文を最大5000例抽出
- そのうち100例をタグ付けの対象とする

読みの付与

- 正解の読みを人手で付与

- 大学院の学生(2名)が担当
- 各自が25単語×100=2500用例の読みの付与を担当
- 作業者の一致率は測っていない 🤖

- 後処理

- 白井が担当
- 誤りチェック
- どの読みも最低10個の用例を含むように用例を入れ換え
 - (標本抽出データセットとはいえ)読みの用例があまりにも少ないのは望ましくない

標本抽出データセット

単語	読み	最小	内訳
一味	2	12	イチミ(88) ヒトアジ(12)
一目	2	16	ヒトメ(84) イチモク(16)
一見	2	10	イッケン(90) イチゲン(10)
一途	2	26	イチズ(74) イット(26)
上手	3	16	ジョウズ(67) カミテ(17) ウワテ(16)
人気	2	10	ニンキ(90) ヒトケ(10)
市場	2	15	シジョウ(85) イチバ(15)
心中	2	37	シンジュウ(63) シンチュウ(37)
寒気	2	13	カンキ(87) サムケ(13)
目下	2	25	モッカ(75) メシタ(25)
色紙	2	12	シキシ(88) イロガミ(12)
出所	2	36	シュッショ(64) デドコロ(36)
方々	2	17	カタガタ(83) ホウボウ(17)
止める	2	40	トメル(60) ヤメル(40)
汚れる	2	15	ヨゴレル(85) ケガレル(15)
素性	2	26	スジョウ(74) ソセイ(26)
素振り	2	42	スブリ(58) ソブリ(42)

単語	読み	最小	内訳
脅かす	2	10	オビヤカス(90) オドカス(10)
辛い	2	45	カライ(55) ツライ(45)
開く	2	10	ヒラク(90) アク(10)
表	2	21	ヒョウ(79) オモテ(21)
角	3	17	カク(52) ツノ(31) カド(17)
上方	2	41	カミガタ(59) ジョウホウ(41)
金星	2	19	キンセイ(81) キンボシ(19)
仮名	2	30	カナ(70) カメイ(30)
一言	2	10	ヒトコト(90) イチゴン(10)
足跡	2	25	アシアト(75) ソクセキ(25)
今日	2	21	コンニチ(79) キョウ(21)
一端	2	10	イッタン(90) イッパシ(10)
大家	2	38	タイカ(62) オオヤ(38)
一角	2	10	イッカク(90) ヒトカド(10)
一行	2	11	イッコウ(89) イチギョウ(11)
根本	2	10	コンポン(90) ネモト(10)
大勢	2	10	オオゼイ(90) タイセイ(10)

単語	読み	最小	内訳
半月	2	44	ハンツキ(56) ハンゲツ(44)
黒子	3	10	クロコ(74) ホクロ(16) コクシ(10)
風車	2	12	フウシャ(88) カザグルマ(12)
背筋	2	32	ハイキン(68) セスジ(32)
生花	2	19	セイカ(81) イケバナ(19)
一寸	2	34	イッスン(66) チョット(34)
見物	2	10	ケンブツ(90) ミモノ(10)
弾く	2	15	ヒク(85) ハジク(15)
御供	2	42	オトモ(58) ゴクウ(42)
細目	2	20	サイモク(80) ホソメ(20)
頭数	2	10	トウスウ(90) アタマカズ(10)
風穴	2	20	フウケツ(80) カザアナ(20)
翡翠	2	10	ヒスイ(90) カワセミ(10)
最中	2	10	サイチュウ(90) モナカ(10)
細々	2	13	ホソボソ(87) コマゴマ(13)
物心	2	41	モノゴコロ(59) ブッシン(41)

標本抽出データセット

単語	読み	最小	内訳
一味	2	12	イチミ(88) ヒトアジ(12)
一目	2	16	ヒトメ(84) イチモク(16)
一見	2	10	イッケン(90) イチゲン(10)
一途	2	26	イチズ(74) イット(26)
上手	3	16	ジョウズ(67) カミテ(17) ウワテ(16)
人気	2	10	ニンキ(90) ヒトケ(10)
市場	2	15	シジョウ(85) イチバ(15)
心中	2	37	シンジュウ(63) シンチュウ(37)
寒気	2	13	カンキ(87) サムケ(13)
目下	2	25	モッカ(75) メシタ(25)
色紙	2	12	シキシ(88) イロガミ(12)
出所	2	36	シュッショ(64) デドコロ(36)
方々	2	17	カタガタ(83) ホウボウ(17)
止める	2	40	トメル(60) ヤメル(40)
汚れる	2	15	ヨゴレル(85) ケガレル(15)
素性	2	26	スジョウ(74) ソセイ(26)
素振り	2	42	スブリ(58) ソブリ(42)

単語	読み	最小	内訳
脅かす	2	10	オビヤカス(90) オドカス(10)
辛い	2	45	カライ(55) ツライ(45)
開く	2	10	ヒラク(90) アク(10)
表	2	21	ヒョウ(79) オモテ(21)
角	3	17	カク(52) ツノ(31) カド(17)
上方	2	41	カミガタ(59) ジョウホウ(41)
金星	2	19	キンセイ(81) キンボシ(19)
仮名	2	30	カナ(70) カメイ(30)
一言	2	10	ヒトコト(90) イチゴン(10)
足跡	2	25	アシアト(75) ソクセキ(25)
今日	2	21	コンニチ(79) キョウ(21)
一端	2	10	イッタン(90) イッパシ(10)
大家	2	38	タイカ(62) オオヤ(38)
一角	2	10	イッカク(90) ヒトカド(10)
一行	2	11	イッコウ(89) イチギョウ(11)
根本	2	10	コンポン(90) ネモト(10)
大勢	2	10	オオゼイ(90) タイセイ(10)

単語	読み	最小	内訳
半月	2	44	ハンツキ(56) ハンゲツ(44)
黒子	3	10	クロコ(74) ホクロ(16) コクシ(10)
風車	2	12	フウシャ(88) カザグルマ(12)
背筋	2	32	ハイキン(68) セスジ(32)
生花	2	19	セイカ(81) イケバナ(19)
一寸	2	34	イッスン(66) チョット(34)
見物	2	10	ケンブツ(90) ミモノ(10)
弾く	2	15	ヒク(85) ハジク(15)
御供	2	42	オトモ(58) ゴクウ(42)
細目	2	20	サイモク(80) ホソメ(20)
頭数	2	10	トウスウ(90) アタマカズ(10)
風穴	2	20	フウケツ(80) カザアナ(20)
翡翠	2	10	ヒスイ(90) カワセミ(10)
最中	2	10	サイチュウ(90) モナカ(10)
細々	2	13	ホソボソ(87) コマゴマ(13)
物心	2	41	モノゴコロ(59) ブッシン(41)

最小の読みが10回

標本均衡化データセットの構築

- 構築手順
 - 低頻度の読みの用例を追加する
 - 目標: どの読みも最低30件の用例をデータセットに含める
- どのように用例を追加するか?
 - ×日本語Wikipediaから抽出した文(最大5000件)から低頻度の読みの用例を検索して追加する
 - そもそも低頻度の用例を見つけるのが困難
 - ✓大規模言語モデル(LLM)に低頻度の読みの用例を生成させる

LLMによる低頻度の読みの用例の生成

- 使用したLLM
 - ChatGPT (GPT 3.5)
- プロンプトの検討
 - 指定された(低頻度の)読みを持つ用例を作る
 - 30個の用例を作る
 - 50文字以上の用例を作る
 - なるべく多様な用例のセットを作る
- 単純なプロンプトではうまくいかない
 - 一味(ひとあじ)を含む文をいくつか作文して下さい。

プロンプト検討例 一見(イチゲン) (1/6)

- 1回目の検討 — 失敗
- 定義文を与える

一見の例文を30個生成してください。

この一見はいちげんという読み方である。この単語の意味は以下の通りです。

(「げん」は「げんざん(見参)」の略) 初めて対面すること。

なじみでなく、初めてであること。

もと、上方の遊里で「初会」の意に用いたが、のち、一般町家でも用いた。一面識。

婚礼諸式の一つ。婿が初めて嫁の親に対面すること。

初婿入り。もともと婿入婚の儀式であったが、嫁入婚になって意味内容が混乱し、親族の初対面や里帰りの挨拶の意にまで使われるようになった。

一見には一見さんという使い方もある。一見さんの意味は初めてそのお店に来店されたお客様の総称である。

一見はいつけんとも読むがこちらは除外する。

いつけんの意味は以下の通りです。

一度見ること。一通り見ること。

ちらっと見ること。(副詞的に用いて) ちょっと見ること。

プロンプト検討例 一見(イチゲン) (2/6)

- 2回目の検討 ― 失敗
- 用例を生成する際のルールを指定する

お願い :イチゲンという読みをもつ「一見」または「一見さん」という単語を含んだ例文を30個以上生成してください。

ルールは必ず守らなければならない。

情報 : この単語の意味 は以下の通りです。

・(「げん」は「げんざん(見参)」の略) 初めて対面すること。

婚礼諸式の一つ。婿が初めて嫁の親に対面すること。

「昨日の客は、一見の田舎の侍だった。」

・一見には一見さんという使い方もある。一見さんの意味は飲食店において初めてそのお店に来店されたお客様の総称である。

得意客ではなく、店の関係者の誰とも面識を持たない客や得意客の紹介を持たない客などを指す。

「あの高級旅館に一度は泊ってみたいものだが、一見さんお断りというのは本当だろうか。」

ルール :

・例文の長さは必ず50文字以上である。

・以下の単語で「一見」と「一見さん」を修飾してはならない。

「初めて」、「初対面」

・一見が以下の意味を持つ場合、いっけんと読まれる。いっけんと読む一見の文章は作成してはいけない。

一度見ること。ひととおりの目を通すこと。「一見に値する」「百聞は一見に如(し)かず」

ちらっと見ること。「一見して事の重大さを悟った」

(副詞的に用いて) ちょっと見たところ。「一見まじめそうな人」

プロンプト検討例 一見(イチゲン) (3/6)

- 3回目の検討 — 失敗
 - ルールを箇条書きにする

##いちげんと読む「一見」または「一見さん」という単語を含んだ例文を生成してください。

1 から 9 のルールをすべて守ってください。

1. 文章には必ず多様性が必要である。
2. 文章は30個以上生成する。
3. 1つの文章の長さは50文字以上である。
4. 「一見」は以下の意味を持つ。

初めて対面すること。

婚礼諸式の一つ。婿が初めて嫁の親に対面すること。

例文：「昨日の客は、一見の田舎の侍だった。」

5. 「一見さん」の以下の意味を持つ。

飲食店において初めてそのお店に来店されたお客様の総称である。得意客ではなく、店の関係者の誰とも面識を持たない客や得意客の紹介を持たない客などを指す。

例文：「あの高級旅館に一度は泊ってみたいものだが、一見さんお断りというのは本当だろうか。」

6. 「一見」を含む文章は10個以上生成する。
7. 「一見さん」を含む文章は10個以上生成する。
8. 文章内で「一見」または「一見さん」と意味が重複する単語を使ってはならない。
9. いっけんと読む「一見」の文章は生成しない。いっけんは以下の意味を持つ。

一度見ること。ひととおりの目を通すこと。

ちらっと見ること。

(副詞的に用いて) ちょっと見たところ。

プロンプト検討例 一見(イチゲン) (4/6)

- 4回目の検討 — 失敗
 - プロンプトを何回かに分け、最初に定義文と例文を与える

1回目

いちげんと読む「一見」は以下の意味を持つ。

(「げん」は「げんざん(見参)」の略) 初めて対面すること。初めて会うこと。特に、旅館や料理屋などの客がなじみでなく、初めてであること。

婚礼諸式の一つ。婿が初めて嫁の親に対面すること。

例文：「昨日の客は、一見の田舎の侍だった。」

いちげんと読む一見は「一見さん」と使われることもある。以下の意味を持つ。

飲食店において初めてそのお店に来店されたお客様の総称である。得意客ではなく、店の関係者の誰とも面識を持たない客や得意客の紹介を持たない客などを指す。

例文：「あの高級旅館に一度は泊ってみたいものだが、一見さんお断りというのは本当だろうか。」

2回目

いちげんと読む「一見」を使った例文を30個生成して。

3回目

すべての例文の長さを50文字以上にして。

プロンプト検討例 一見(イチゲン) (5/6)

- 5回目の検討 — 失敗
 - 意味の定義を箇条書きで与える

1回目

「一見」をいちげんと読む場合以下の意味を持つ。

1. (「げん」は「げんざん (見参)」の略) 初めて対面すること。
 2. なじみでなく、初めてであること。特に、旅館や料理屋などの客がなじみでなく、初めてであること。
- 2の例文は「昨日の客は、一見の田舎の侍だった。」
3. 婚礼諸式の一つ。婿が初めて嫁の親に対面すること。初婿入り。もともと婿入婚の儀式であったが、嫁入婚になって意味内容が混乱し、親族の初対面や里帰りの挨拶の意にまで使われるようになった。
 4. 「一見」をいちげんと読む場合「一見さん」と使われることもある。これは以下の意味を持つ。
飲食店において初めてそのお店に来店されたお客様の総称である。得意客ではなく、店の関係者の誰とも面識を持たない客や得意客の紹介を持たない客などを指す。
- 4の例文「あの高級旅館に一度は泊ってみたいものだが、一見さんお断りというのは本当だろうか。」

2回目

これらの例文を作成しろ。

3回目

以下の全てのルールに従ってこれらの例文を作成しろ。

1. 文章を30個以上生成する。
2. 文章の長さは200文字以上である。

プロンプト検討例 一見(イチゲン) (6/6)

- 6回目の検討 — 成功！
- 語義毎に用例を生成するように指示

1回目

「一見」をいちげんと読む場合以下の意味を持つ。

1. (「げん」は「げんざん (見参)」の略) 初めて対面すること。
 2. なじみでなく、初めてであること。特に、旅館や料理屋などの客がなじみでなく、初めてであること。
- 2の例文は「昨日の客は、一見の田舎の侍だった。」
3. 婚礼諸式の一つ。婿が初めて嫁の親に対面すること。初婿入り。もともと婿入婚の儀式であったが、嫁入婚になって意味内容が混乱し、親族の初対面や里帰りの挨拶の意にまで使われるようになった。
 4. 「一見」をいちげんと読む場合「一見さん」と使われることもある。これは以下の意味を持つ。
飲食店において初めてそのお店に来店されたお客様の総称である。得意客ではなく、店の関係者の誰とも面識を持たない客や得意客の紹介を持たない客などを指す。
- 4の例文は「あの高級旅館に一度は泊ってみたいものだが、一見さんお断りというのは本当だろうか。」

2回目

1 から 4 の例文を8個ずつ作成しろ。

3回目

それぞれの文章を50文字以上にしろ。

ChatGPTで生成した用例の例

● 一見(イチゲン)

音楽フェスでの*一見*の瞬間に、未知のアーティストに魅了されました。
*一見*の出会いが、新たな仕事上の結びつきと友情を築き上げました。
*一見*のデートで、彼女の笑顔に心が躍り、初対面ながら、彼女の魅力に引き込まれ、未知の感情が芽生えました。
交流会の*一見*の瞬間に、ビジネスパートナーシップが始まりました。
展示会での*一見*の出会いが、アートとビジネスの融合を生み出しました。
街での*一見*の出会いが、後に大規模なプロジェクトに繋がりました。
地元のフェスティバルには、*一見*の参加者が全国から訪れました。
*一見*の訪問者に、地元の特産品が贈られる地域イベントが開催されました。
町の新しい公園には、*一見*の家族連れが笑顔で訪れていました。
*一見*の来客に伝統的な文化体験が提供されるツアーが企画されました。
地元の*一見*のイベントには、地元民として初めて参加する喜びがありました。
新しく開かれたギャラリーには、*一見*のアート愛好者が集まっていました。
この美容室は、*一見*さんでも気軽に立ち寄れるアットホームな雰囲気があります。
車の修理は初めてだが、*一見*さんでもわかりやすい説明がありました。
*一見*さん向けの料理教室が地元のコミュニティセンターで開催されました。
*一見*さんにも優しいスタッフが新しい図書館で読書の楽しさを案内してくれました。
この公園は*一見*さん歓迎で、地元の歴史について学ぶことができます。
*一見*さんでも楽しめるツアーが市内の観光協会から提供されています。
*一見*さん向けのアートクラスが地元のアートスタジオで開催されます。
ファッションショーは*一見*さんでも楽しめるように、解説付きで進行されました。

ChatGPTで生成した用例の例 (その2)

● 人気(ヒトケ)

この山間の小道は*人気*のない場所で、ただ自然の静けさが広がっている。
雨の日になると、公園は*人気*が薄れ、静かな雰囲気包まれます。
街の片隅にある小さなカフェは*人気*にとぼしく、ひっそりと営業している。
その廃校はもはや*人気*がなく、草が生い茂る中、寂しさが漂っている。
草原に立つ一軒の家は*人気*のない土地に建てられており、周りに誰もいない。
*人気*が薄れた港町は、かつての繁栄とは対照的に寂れた様子を見せている。
森の奥深くにたどり着くと、*人気*のない秘密の庭園が広がっている。
街の中心部から離れると、道路はますます*人気*がなくなり、静かな風景が広がる。
田舎の小さな村は*人気*がとても薄く、のどかな時間が流れている。
昼間でもこのエリアは*人気*がなく、街のどこかで見かけることはめったにない。
この季節になると、山の中腹は*人気*がなく、登山者もまばらだ。
*人気*が薄れた広場には、草が生い茂り、静けさが漂っている。
その美しい湖は訪れる者が少なく、*人気*があまりない秘境だ。
雪が積もると、街の通りは*人気*が減り、静かで穏やかな雰囲気に包まれる。
昼下がりの公園は*人気*がなく、ベンチに座る人もほとんどいない。
季節外れのビーチは*人気*がなく、静かな波音が響いている。
*人気*が薄れた古い神社は、参拝者の姿もまばらで、静寂に包まれた神聖な雰囲気を醸し出している。
夜になると、街の中心街は*人気*が落ち着き、静かな雰囲気に変わる。
この小道は*人気*がなく、ただ木々の間から差し込む陽光が美しい。
この孤立した村は*人気*にとぼしく、他の地域とは異なる雰囲気が漂っている。

プロンプトのまとめ (1/2)

- 指定された読みの用例を生成させるには...
 - 語義の定義文を示すとよい
 - 例をいくつか与えるとよい
 - (語義が複数あるとき) 語義毎の用例を作文するよう指示する方がよい
- 段階に分けた方がよい
 1. 語義の定義と例文を提示
 2. 用例の生成を依頼
 3. 生成した用例の文長を長くすることを依頼

プロンプトのまとめ (2 / 2)

● 問題点

- 文が短い(文字数を指定しても複数の文を生成する)
 - 音楽フェスでの*一見*の瞬間に、未知のアーティストに魅了されました。まだ知らない才能に触れ、新しい音楽の世界へのドアが開かれ、私の音楽体験が豊かになりました。その瞬間は、感動と発見の絶妙な融合でした。
- 誤りがある(正しい読みの文が生成できていない)
 - この大切な場で*一言*挨拶を申し上げます。
 - 「イチゴン」の読みの例文を生成したかった
- 多様性にやや欠ける

この山間の小道は*人気*のない場所で、ただ自然の静けさが広がっている。

街の片隅にある小さなカフェは*人気*にとぼしく、ひっそりと営業している。

田舎の小さな村は*人気*がとても薄く、のどかな時間が流れている。

標本均等化データセット

単語	読み	合計	最小	内訳
一味	2	117	30	イチミ(87) ヒトアジ(30<17>)
一目	2	114	30	ヒトメ(84) イチモク(30<14>)
一見	2	120	30	イッケン(90) イチゲン(30<20>)
一途	2	104	30	イチズ(74) イット(30<4>)
上手	3	127	30	ジョウズ(67) カミテ(30<13>) ウワテ(30<14>)
人気	2	120	30	ニンキ(90) ヒトケ(30<24>)
市場	2	115	30	シジョウ(85) イチバ(30<15>)
心中	2	100	37	シンジュウ(63) シンチュウ(37)
寒気	2	117	30	カンキ(87) サムケ(30<17>)
目下	2	105	30	モッカ(75) メシタ(30<5>)
色紙	2	118	30	シキシ(88) イロガミ(30<18>)
出所	2	100	36	シュッショ(64) デドコロ(36)
方々	2	113	30	カタガタ(83) ホウボウ(30<13>)
止める	2	100	40	トメル(60) ヤメル(40)
汚れる	2	115	30	ヨゴレル(85) ケガレル(30<15>)
素性	2	104	30	スジョウ(74) ソセイ(30<4>)
素振り	2	100	42	スブリ(58) ソブリ(42)
脅かす	2	120	30	オビヤカス(90) オドカス(30<20>)
辛い	2	100	45	カライ(55) ツライ(45)
開く	2	120	30	ヒラク(90) アク(30<20>)
表	2	109	30	ヒョウ(79) オモテ(30<9>)
角	3	113	30	カク(52) ツノ(31) カド(30<13>)
上方	2	100	41	カミガタ(59) ジョウホウ(41)
金星	2	111	30	キンセイ(81) キンボシ(30<11>)
仮名	2	100	30	カナ(70) カメイ(30)

単語	読み	合計	最小	内訳
一言	2	120	30	ヒトコト(90) イチゴン(30<20>)
足跡	2	105	30	アシアト(75) ソクセキ(30<5>)
今日	2	109	30	コンニチ(79) キョウ(30<9>)
一端	2	120	30	イッタン(90) イッパシ(30<25>)
大家	2	100	38	タイカ(62) オオヤ(38)
一角	2	120	30	イッカク(90) ヒトカド(30<20>)
一行	2	119	30	イッコウ(89) イチギョウ(30<19>)
根本	2	120	30	コンポン(90) ネモト(30<20>)
大勢	2	120	30	オオゼイ(90) タイセイ(30<20>)
半月	2	100	44	ハンツキ(56) ハンゲツ(44)
黒子	3	134	30	クロコ(74) コクシ(30<20>) ホクロ(30<14>)
風車	2	118	30	フウシャ(88) カザグルマ(30<18>)
背筋	2	100	32	ハイキン(68) セスジ(32)
生花	2	111	30	セイカ(81) イケバナ(30<11>)
一寸	2	100	34	イッスン(66) チョット(34)
見物	2	120	30	ケンブツ(90) ミモノ(30<20>)
弾く	2	115	30	ヒク(85) ハジク(30<15>)
御供	2	100	42	オトモ(58) ゴクウ(42)
細目	2	110	30	サイモク(80) ホソメ(30<10>)
頭数	2	120	30	トウスウ(90) アタマカズ(30<20>)
風穴	2	110	30	フウケツ(80) カザアナ(30<10>)
翡翠	2	120	30	ヒスイ(90) カワセミ(30<21>)
最中	2	120	30	サイチュウ(90) モナカ(30<20>)
細々	2	117	30	ホソボソ(87) コマゴマ(30<17>)
物心	2	100	41	モノゴコロ(59) ブッシン(41)

< >はChatGPTで生成した用例数

おわりに

- 日本語読み推定タスク・データセットを構築しました
 - 50単語 $\times(100 + \alpha)$ の用例に対して対象単語の正解の読みを付与したデータセット
 - 用例は Wikipediaの文 + ChatGPTで生成させた文
 - 標本抽出データセット、標本均等化データセット
- 近日中に公開する予定です