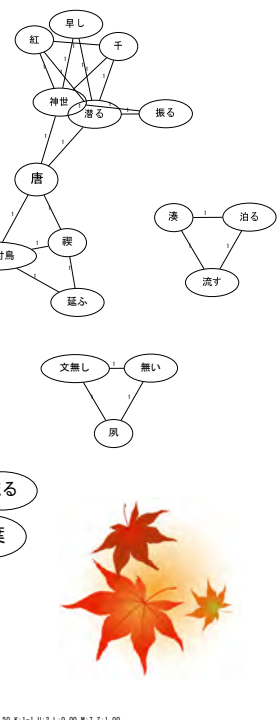


二十一代集シソーラスのための漸近的語彙対応システムの開発

山元啓史†* 村井源† ホドシチェクボル‡
 †東京工業大学 *カリフォルニア大学サンディエゴ校 ‡大阪大学

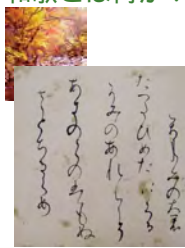


目的

- プロジェクトの目的は、和歌用語のシソーラスを開発することである。
- 本研究では、そのための語彙の自動対応づけシステムの開発を行う。
- 和歌とその現代語訳 10 種を用いてパラレルコーパスを作成し、対応づけに用いる。
- 将来的には、これを利用して、客観的な基準による語彙の分類と言語変化の探究を行う。

問題

和歌とは何か？



立田姫
 手向くる神の
 あればこそ
 秋の木の葉の
 幣と散るらめ

1. 表記の問題

龍田, 立田, 竜田, たつた

2. 単位の問題

卯の花? 卯/の/花?

3. 所属の問題

「卯の花」は植物か食品か?

4. 掛詞の問題

「海松藻(みるめ)」と「見る目」?



方法

材料: 古今和歌集 (905 年頃) とその現代語訳 10 種

Mutual Co-occurrence Rate: 村井 (2010)

$$mcr(o, t) = p(o | t) p(t | o)$$

ただし、 o はオリジナル (和歌) の単語。 t は現代語訳の単語。 $mcr(o, t)$ は相互共出現率。 $p(o | t)$ は、オリジナルと現代語訳の対応する 2 つの文に注目した時、単語 o と単語 t が同時に出現した割合。
 → $mcr(o, t)$ の値が、十分に大きい時、単語 o と単語 t は、**文脈的に一致している**と推定。

結果

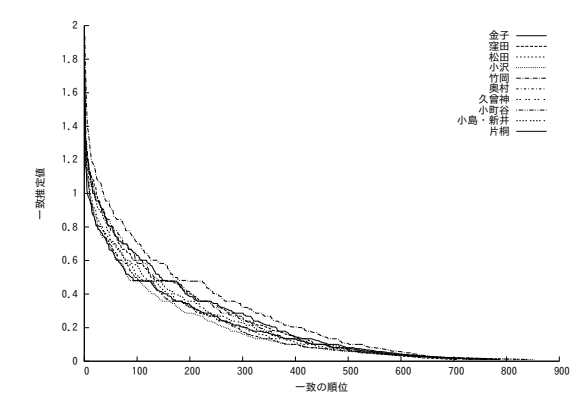


図 2: 古今集と現代語訳 10 種中の任意の単語対の mcr 値の分布。

表 1: 「古今和歌集」とその現代語訳 10 種の単語の対応処理を実施し、その推定値の上位、中位、下位ごとの単語の対応例

no.	上位対 (1.3 以上)	中位対 (0.16 より)	下位対 (0.01 以下)
1	鳴く 鳴く	老ゆ 年老いる	異なり あの
2	風 風	乱る 乱れる	雫 どうして
3	世の中 世の中	来 いらっしやる	此の この
4	人 人	問ふ 問う	隨に まま
5	春 春	問ふ 訪ねる	匂ふ 美しい
6	秋 秋	名 噂	見る せい
7	時鳥 時鳥	変はる 変る	連れ つく
8	時鳥 ほととぎす	燃ゆ 燃える	立ち返る 言う
9	散る 散る	濡づ 濡れる	有り つく
10	見る 見る	難し むずかしい	有り まさしく

考察

- 動詞: 「落つ 落ちる (自動詞)」, 「捨つ 捨てる (他動詞)」
- 名詞: 「木綿付鳥 (ゆふつけどり) 鶏」, 「朝な朝な 毎朝」
- 古語固有: 機械的に、古語「名」に対して「噂」が推定できた。
- 推定範囲: 単語対の推定が可能なのは 0.2 あたりまで。

おわりに

和歌・現代語訳のパラレルコーパスを利用した単語対の推定法により、新規シソーラスに追加すべき語の決定はほぼ計算手続きによってできるだけでなく、比喻、含意、掛詞などの和歌特有の対応づけも可能なことがわかった。

文献

● 村井源 (2010) 「漸近的対応語彙推定法に基づく翻訳文の解釈的特徴の抽出: 日本語翻訳聖書の計量的比較」, 『情報知識学会誌』, 第 20 巻, 第 3 号, 293-310.