

品詞による文評価を用いた日本語語呂自動生成手法

岡安 優弥 高田 雅倫 渡辺 邦浩 濱川 礼

中京大学 情報理工学部 情報システム工学科

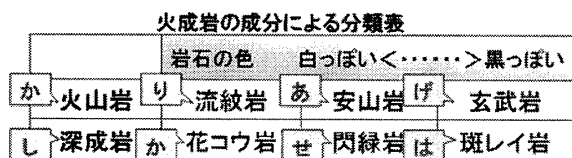
1. 概要

本論文では日本語語呂合わせ（以下、語呂とする）の自動生成手法について述べる。本研究における語呂とは、複数の単語が規則性を持って並んでいる順番を覚える際に使われるものとする。良い語呂の定義は、覚えやすく、元の単語が思い出しやすい語呂である。本研究では、より良い語呂を生成するための手法を提案し、システムとして実装、実際に良い語呂が生成できるかの検証を行った。

2. 背景・目的

暗記学習において、語呂を用いることは、単語列や数字列の順序を正確に覚える際に非常に有効な手段であり、広く用いられている。

図 1 は、中学校の学習範囲の火成岩の種類の成分別に並べた順番を覚える際に、実際に使われている語呂の例である。覚えたい単語の頭文字を抽出し、抽出した要素を持った、日本語文章を作成している。本研究では、この語呂生成の過程の自動化できる手法の提案を目的としている。



刈り上げの新幹線は速い

図 1 語呂合わせ実例

3. 提案手法

提案手法では、暗記対象となる要素からそれぞれの頭文字を抜き出して、それを含んだ単語に置き換えることで語呂の生成を行う。実装したシステムでの処理手順を図 2 に示す。

最初に頭文字の抜き出し、データベースより置き換え単語の検索、検索結果として得られた候補単語を品詞に応じて刈り込み、残った候補単語の評価値算出、単語間への助詞の挿入という流れで語呂を生成する。

Japanese pun automatic generation method
that uses sentence evaluation by part of
speech

Yuya Okayasu, Masanori Takada,
Kunihiro Watanabe and Rei Hamakawa

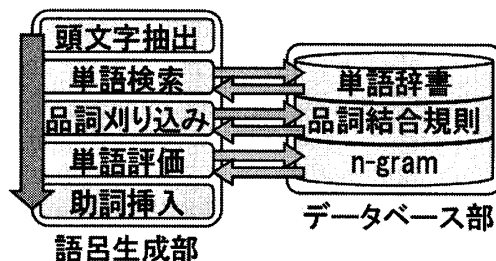


図 2 処理の手順

置き換え候補単語は、検索クエリにワイルドカードを含めた緩い条件での検索を行い、複数の候補を取得する。該当する語がない場合は、クエリを一文字短くして該当しやすくし再検索する。尚、今回実装したシステムでは、辞書データベースとしてオープンソース形態素解析エンジン MeCab[1]の辞書を加工して使用した。単語検査の例を図 3 に示す。

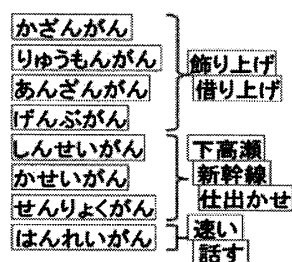


図 3 置き換え単語検索

置き換え単語検索の後、以下の 4 つの評価によって実際に語呂生成に使用する単語を決定する。単語が決定した後、品詞結合規則に定義された助詞を挿入して語呂が完成する。

▶ 単語評価 1 品詞結合規則による品詞評価

本研究では品詞の並び方を整えることで、文法的に正しく、文として成立した語呂を生成することとした。

文法的に正しい品詞の組み合わせをデータベースとして複数定義しておき、候補単語に応じて使用可能なものをデータベースより選択することで、文法的に正しい品詞の並びの語呂を生成することができる。実装したシステムでは、図 4 の例に示すような、文として成立する品詞の並び方を品詞結合規則として 18,657 通り定めた。品詞結合規則に適合していない単語は候補から除外し、語呂生成に使用しない。

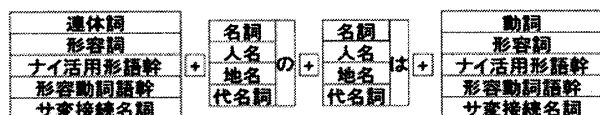


図 4 品詞結合規則の例

➤ 単語評価 2 n-gram を用いた単語の評価

n-gram とは、任意の範囲の文章中で単語や単語列の出現する割合を数値で示したデータ [2] である。n-gram の値が高い単語と単語列を優先して用いることで、より使われやすい単語や単語列を語呂に含ませるができる。本研究ではこの性質を利用することで、単語間に関連を持たせ、語呂全体にまとまりを持たせている。

実装したシステムの n-gram データには、Google が開発し言語資源協会が提供している「Web 日本語 N グラム第 1 版」を利用している。

➤ 単語評価 3 頭文字以降の文字の一致評価

語呂は一般的に頭文字のみを用いて作成されるため、語頭が同じ要素が複数あると識別が難しいという問題がある。そこで本研究では、入力要素の中に頭文字が重複する単語がある場合、入力要素と検索語について頭文字以降の文字が一致するかを判定して評価を行う。

例として[かざんがん]の頭文字[か]を用いて単語検索を行い、[かりあげ]と「かざりあげ」が検索された場合、図 5 に示すように二文字目一致を判定して[かざりあげ]を優先的に扱う。

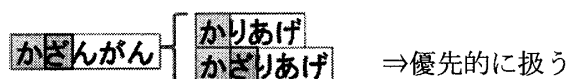


図 5 二文字目以降の一致

➤ 単語評価 4 雑音率による単語の評価

語呂において、元要素の頭文字以外の文字は、適当な単語を当てはめるために挿入した雑音であるため、語呂から元の要素を思い出す復号時に邪魔になる。

そこで本研究では、単語に含まれている雑音文字の数に応じた減点評価を行うことで、語呂中の雑音が少なくなるようにしている。図 6 に、赤枠で囲われた文字が雑音である場合の雑音率の例を示す。

$$\frac{\text{雑音文字数}}{\text{全文字数}} = \frac{2}{5} \dots \text{雑音率 } 40\%$$

$$\frac{\text{雑音文字数}}{\text{全文字数}} = \frac{1}{4} \dots \text{雑音率 } 25\%$$

⇒優先的に扱う

図 6 雑音率の決定

尚、提案手法 3 で示した二文字目以降の一致評価に該当した文字は、雑音から除外する。

品詞結合規則による品詞評価の後、前述の 3 つの評価値に重みを掛けて足し合わせた総合評価値によって最終的な使用単語を決定する。評価値の重みは、それぞれの評価が同程度のばらつきを生むように設定した。

使用語が決定した後、品詞に応じた助詞を単語間に挿入して語呂が完成する。

4. 成果

本研究における提案手法を実装したシステムを用いて生成された語呂の例を示す。

- 地球科学 地質時代の区分(4 要素)
 - ◇ 顕生代, 原生代, 始生代, 冥王代
 - ⇒**顕現する使命**
- 生物 被子植物種子の多い上位(5 要素)
 - ◇ 菊, 蘭, 豆, 稲, 茜
 - ⇒**気楽なマイアミ**
- 歴史 近代までのオランダの歴史(6 要素)
 - ◇ フランク王国, 神聖ローマ帝国, ブルゴーニュ公爵領, ネーデルラント 17 州, ハプスブルク領, 連邦共和国
 - ⇒**不思議なブルネイは破裂する**
- 心理学 エリクソン発達課題 (8 要素)
 - ◇ 不信, 恥, 劣等感, 拡散, 孤独感, 自己吸収, 嫌悪
 - ⇒**不発は劣化する好条件**

また、実装したシステムを用いて実際に語呂を生成し評価を行った。生成元要素数の異なる 3 種類の語呂を評価対象者がそれぞれ 2 つずつ、各自が選択した暗記対象の要素を用いて生成して評価を行った。結果、上記の生成例のような短い語呂あわせに対して好評価を得られた。

5. 考察

上記に示すことから、覚えやすい語呂の生成に有用な手法を提案することができた。

しかし、長い語呂を生成する場合の評価が芳しくなかったことから、長大な語呂の生成方法に改善の余地があることがわかる。

今後、より覚えやすい語呂の自動生成を行う際には、長大な語呂を如何に短く、覚えやすくするかを考慮することが重要である。

参考文献

[1] オープンソース 形態素解析エンジン MeCab
<http://mecab.sourceforge.net/>
 [2] n-gram モデルを利用したテキスト分析
<http://www.shuiren.org/chuden/teach/n-gram/index-j.html>