3G3-OS-05a-6

Bi-gram 連接表と単語列変形規則に基づく回文自動生成

Automatic Japanese Palindrome Generation Based on Phrase Bi-grams and Transformation Rules

清野 舜 Shun Kiyono　渡邉 研斗 Kento Watanabe　岡崎 直観 Naoaki Okazaki　乾 健太郎 Kentaro Inui

東北大学
Tohoku University

Palindrome generation requires its output not only to fulfill the reading condition of palindrome but also to be a meaningful sequence of words. In addition, palindromes often exhibit the idiosyncratic structure and style of language; thus, a language model simply induced from an existing text corpus may not sufficiently cover the potential variety of palindromes. Addressing those issues, this paper proposes a new method for generating Japanese palindromes based on phrase bi-grams and transformation rules. Our experiment empirically shows that our method considerably outperforms an existing baseline method in producing meaningful palindromes out of a much smaller number of candidates.

1. はじめに

回文とは、文頭・文末から読んだ場合で、音節の順序が変わらず、かつ言語として無理なく意味が通る文であり、言葉遊びの一種である。本論文では、回文を次の 2 つの条件を満たす文であると定義する。(1) **音の条件**：文頭から読んでも文末から読んでも同じ音であること、(2) **通意条件**：日本語として無理なく意味が通じること。計算機の場合は、音の条件を満たす文字列の生成だけなら無限に生成可能であるが、その中から通意条件を満たすものを選ぶ必要がある。一方、人間の場合は「雪の下 待ちわびた皆草 桜が楽さ 作並 旅は巷 しのぎ湯」-「ゆきのした まちわびたみなくさ さくらがらくさ さくなみ たびはちまた しのぎゆ」[*1] のような、両条件を満たした非常に長い回文を作成することができる。本研究では、人間が作ったような通意条件を満たした回文の自動生成に挑戦する。

既存手法における回文生成手法では、通意条件を考慮せず音の条件のみを満たすように、文節の組み合わせを総当りで探索している。そのため、意味の通じる回文の割合は非常に少ないだけでなく、生成に必要とする計算量が膨大になる。本研究では、通意条件を満たす回文を多く生成するために、文節を単位とする bi-gram 連接表を用いて文節間の依存関係を考慮した回文生成手法を提案する。

また、回文では「妻どかす門松」のように、助詞の省略や倒置表現など、通常の日本語文とは異なった文体が多く存在するため、単純な日本語のコーパスから構築された文節の bi-gram では、回文がうまく生成できないことが考えられる。そこで本研究では、回文によく使われる「単語の省略・倒置・置換」の変形規則を作成し、bi-gram の拡張を行った。

本研究では、変形規則を適用する前と後での生成結果を比較し、変形規則が及ぼす生成回文への影響について分析した。特に、助詞の省略処理を適用することで、回文の生成数が大きく増えることが分かった。また、bi-gram を使用した結果、通意条件を満たす回文を、既存手法よりも高い割合で生成することに成功した。

連絡先: 清野舜，東北大学工学部情報知能システム総合学科，宮城県仙台市青葉区荒巻字青葉 6-6-05，022-795-7091，022-795-4285，shun.kiyono@dc.tohoku.ac.jp

*1 第 17 回 日本ことば遊び回文コンテスト入選作品より引用

2. 回文の生成手順

本章では、人間が回文を作る際の手順と、計算機への応用研究について説明する。主な作成手法は以下に述べる**文頭固定法** [1] と**折り返し固定法** [1] の二種類に大別できる。人間による回文の作文は、多くの場合この 2 種 (またはその組み合わせ) が採用されている。計算機の場合も、2 種どちらの手法でも回文生成が可能である。両方の手法とも、文節/単語単位の生成が可能であり、文節/単語全体の集合を探索対象として用いている。

2.1 折り返し固定法

折り返し固定法とは、ある単語から左右に単語を付け加えながら回文を作成する方法である。

折り返し固定法による「妻どかす門松」生成

1. 単語「どかす」の「す」を回文の中央の文字とする
2. 「どかす」の右に「かど」で始まる「門松」を選ぶ
3. 「まつ」がはみ出るため、「妻」を左に付け足す
4. 結果、音の条件を満たす「妻 どかす 門松」が完成

鈴木らは 3 文節の組み合わせを全て列挙して回文を生成した。[1] しかし、約 350 万の文節の組み合わせを用いたため、計算時間は 142.5 日と膨大になり、生成された約 140 万の回文の中で通意条件を満たすものは 49 しかなく、効率は良くない。

本研究では、折り返し固定法の改良に取り組む。

2.2 文頭固定法

文頭固定法は、文頭の単語から音の条件を満たすように単語を伸ばして作成していく方法である。

文頭固定法による「妻どかす門松」生成

1. 文頭「妻」を選ぶ
2. 文末に「まつ」で終わる「門松」を選ぶ
3. 「かど」がはみ出るため、「妻」の後ろに「どか」で始まる「どかす」を選ぶ
4. 結果、音の条件を満たす「妻 どかす 門松」が完成

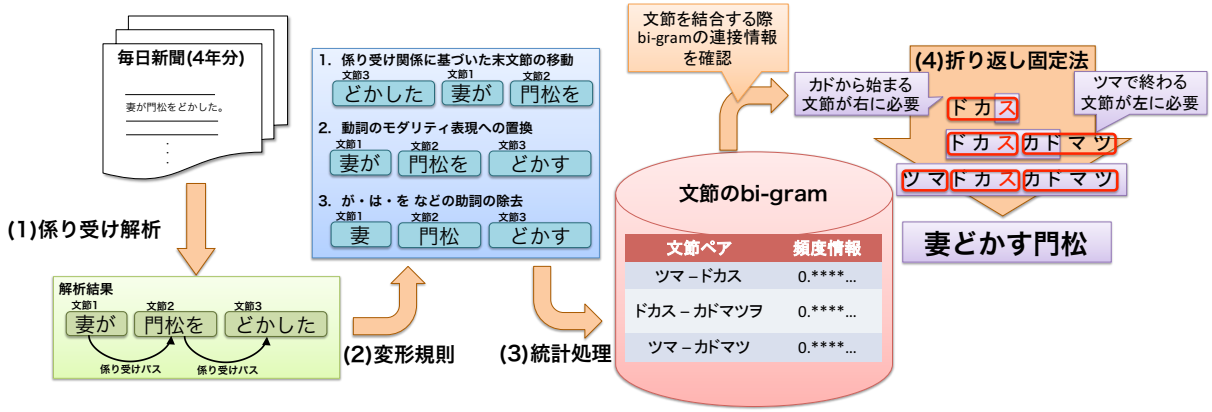

図 1: 回文生成過程 (提案手法) の概要

なお鈴木ら [1] は、この作成手法を用いた回文自動生成を行っている。ここでも鈴木らは文節の組み合わせを全て列挙しており、通意条件を満たす回文の割合は少ない。

2 種類いずれの手法でも、計算機による回文生成は可能である [1, 2] が、既存研究は音の条件を満たした文に対する実際の回文 (通意条件を満たしたもの) の割合が非常に小さい。そこで本論文では、日本語のコーパスから獲得した文節 bi-gram を回文の生成過程に用いることで、文節間の依存関係を考慮し、できるだけ通意条件を満たすような文を探索する手法を提案する。これによって、より通意条件を満たした回文が生成される確率を高めることができる。

上記の例では、単語を付け加えることで回文を作成しているが、文節や文など、大きな単語のまとまりを付け足しても、回文を生成することができる。実際に関連研究 [1, 2] では単語ではなく文節を付け加えることで回文を生成している。

3. 提案手法

本研究では、通意条件を満たす回文を積極的に生成するために、図 1 に示すように、折り返し固定法をベースとして、新たに単語列変形規則と文節の bi-gram を応用した回文生成手法を提案する。提案手法では以下の手順により、通意条件を満たした回文を生成する。

回文生成手順

1. 入力した日本語データ (今回は毎日新聞記事 4 年分) を係り受け解析器 CaboCha を用いて解析し、形態素、文節、係り受け情報を取得
2. 回文特有の文体の文章を生成するために、各文節に変形規則を適用
3. 隣り合う文節ペアの頻度情報から、bi-gram を生成
4. 折り返し固定法を用いて回文を生成する。文節を結合する際は、bi-gram の連接情報を読む (3.1 節参照)
5. 音の条件を満たした文を出力
6. 出力として得られた文の集合から、通意条件を満たすものを人力で抽出

なお、本研究では単語の bi-gram でなく文節の bi-gram を使用している。文節は一つ以上の単語から構成され、一単語より

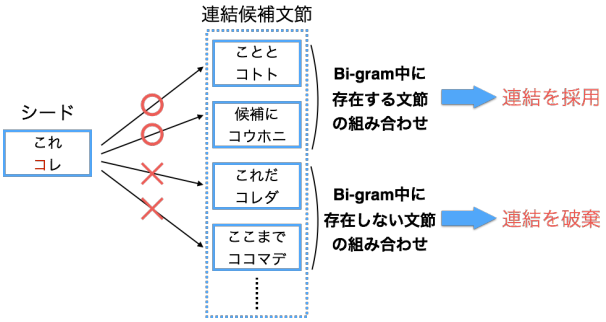

図 2: bi-gram を用いた折り返し固定法

も意味をつかみやすいため、文節同士を結合することで通意条件を満たす回文が生成されると予想した。

本手法では、関連研究 [1, 2] と同様に、折り返し固定法を用いて文節を結合する。以下、3.1 節と 3.2 節、文節の bi-gram を用いた折り返し固定法による回文生成手法と単語列変形規則について、それぞれ詳細を説明する。

3.1 文節の bi-gram を用いた回文生成アルゴリズム

文節の bi-gram を利用した折り返し固定法による回文生成の概要を、図 2 に示した。

任意のシード文節に対して、音の条件を満たすように左右に文節を加えていくという手順は、2.1 節で述べたものと同様である。本研究では、音の条件を満たす文節を新たに連結する際、隣接する文節のペアがコーパス中に一度でも出現していたかを確認し、ペアが存在していた場合に限り連結を実行する。その結果、日本語として不自然な文節列が生成回文に含まれることを防ぐことができる。

3.2 回文に特化した単語列変形規則

前節では通意条件を満たす回文を生成するために、文節の bi-gram を使って回文を生成すると述べたが、その構築ためにはコーパスが必要となる。本研究では、bi-gram の構築のために毎日新聞記事 4 年分のデータを使用した。

しかし、回文には特有の文体が存在するため、単純な bi-gram では表現できない文体が数多く存在する。例えば、「妻どかす門松」 - 「ツマドカスカドマツ」が挙げられる。新聞記事において、「妻どかす門松」のように倒置や省略をすることは殆どなく、「妻が門松をどかした」のようなきれいな文体を使うことが多い。そのため、通常の文章から構築した文節 bi-gram(以

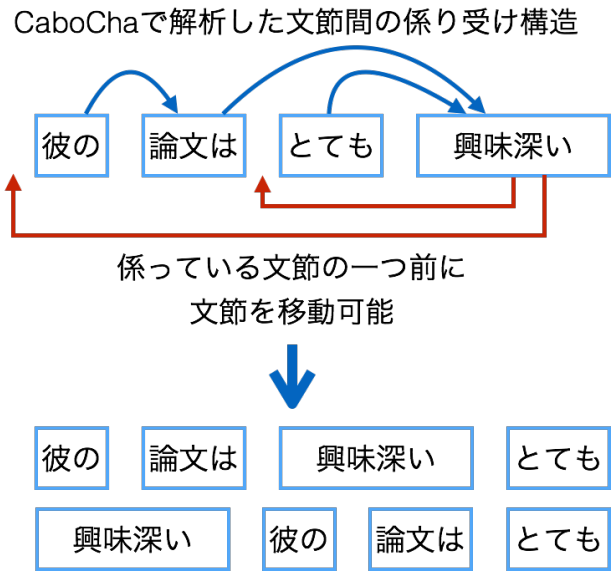

図 3: 係り受けに基づく末文節の移動

下、素 bi-gram と言及) では、「妻どかす門松」のような表現が含まれないため、省略や倒置を用いた回文が生成できない。よって、「妻が門松をどかした」を回文特有の文体に変形することで、素 bi-gram を回文に特化したものに拡張する。本研究では以下に挙げる 3 つの変形規則を作成した。

変形規則

1. 省略処理
   「妻が門松をどかした」→「妻門松どかした」
2. 倒置処理
   「妻が門松をどかした」→「妻がどかした門松を」
3. 置換処理
   「妻が門松をどかした」→「妻が門松をどかす」

本研究では事前処理として、係り受け解析器 CaboCha を用いて、文節間の係り受け構造を解析した。

3.2.1 省略処理

日本語は、主格や目的格を表す格助詞の省略に対して寛容な言語であり、一部の助詞が省略されていても、元の文章の意味を掴めることが多い。例えば「私は野球をする」から助詞「は」と「を」を取り除いた文章「私野球する」でも、意味を理解することが可能である。本研究では助詞「を」「が」「は」を文節から省略したものを bi-gram に追加する。

3.2.2 倒置処理

本研究で取り扱う新聞記事において、「妻どかす門松」のような倒置表現は少ない。そこで新聞記事に、図 3 で示した倒置処理を施した。この処理では、日本語の係り受け構造に着目し、末文節を移動することで文意を損なわないように配慮した。

日本語の係り受け文法では、一般的に次の 3 つの性質に従う: 非交差性（係り受けは互いに交差しない）、後方修飾性（係り先は後方の文節）、係り先の唯一性（各文節は必ず 1 つの文節に係る）。本研究では、非交差性と唯一性を順守しつつ、後方修飾性を破る位置に述語（末文節）を移動することで、新聞記事の文を倒置を用いた文に変形する。

まず CaboCha で解析された文節間の係り受け構造を使い、末文節から係り元の文節を順次たどる。係り元を持たない文節に移動した時 (図 3 上「彼の」「とても」)、その文節の直前に末文節を移動させ、その文節ペアを bi-gram に追加する。これにより、非交差条件を保ったまま、前方修飾性（右から左への係り）を持つ文が完成する。これは、「論文は」の直前に「興味深い」が移動し（非交差性の破壊)、「彼の 興味深い 論文は とても」のような文に変形することを防いでいる。

3.2.3 置換処理

この処理では、動詞の文節「どかした」を「どかす」や「どかしたい」などの様々なモダリティ表現に置換することで、bi-gram 中の動詞を含んだ文節ペアの量を増やす。まず、入力した日本語データの全体を走査し、動詞の活用形＆助動詞＆助詞の組み合わせからなる文節 (モダリティ表現) 全てをその動詞の基本形と紐付けて保存しておく。その後、CaboCha で解析した文章の末文節が動詞の基本形だった場合は、事前に保存しておいたモダリティ表現で置換し、bi-gram に追加する。

以上の処理を加える事で、元の文節の意味を損なうこと無く bi-gram に回文独自の表現を加える事ができる。

表 1: 各モデル毎の回文生成結果

モデル	文節ペアの個数	X:音の条件を満たした文の数	Y:回文の数	Y/X
既存研究 [2]	なし	140 万	49	0.00001
B	10,622,411	41	4	0.098
B+省略	16,430,076	1,126	56	0.050
B+倒置	17,049,175	370	22	0.059
B+置換	214,891,015	112	10	0.089
B+全部	227,153,429	1,773	70	0.039

4. 実験と考察

本研究では、文節の bi-gram を構築するための日本語コーパスとして毎日新聞記事 4 年分 (約 2,000 万文、15 万単語) を用いた。前章で述べた単語列変形規則が回文の生成に及ぼす影響を調べるために、以下の規則を適用した bi-gram を用いて実験を行った。まず、合計 150 万のシード文節からなる集合を用意して、各モデルで回文生成を行ったその後、生成された音の条件を満たした文を読み、通意条件を満たしている文の数を人手で数えた。

- B: 素 bi-gram
- B+省略: 素 bi-gram に省略処理を加えた bi-gram
- B+倒置: 素 bi-gram に倒置処理を加えた bi-gram
- B+置換: 素 bi-gram に置換処理を加えた bi-gram
- B+全部: 素 bi-gram に省略処理、倒置処理と置換処理を加えた bi-gram

表 1 に既存手法 [2] と各モデルが生成した文の数（X)、その中で通意条件を満たした数（Y)、通意条件を満たす割合（Y/X）を示した。さらに、提案手法が用いた文節 bi-gram の数も示した。この表より、「B+全部」に含まれる文節ペア数は「B」より 20 倍以上に増えたことがわかる。また、単語列変形処理を加えたモデルの方が、より多くの回文を生成でき、既存研究より 1,000 倍以上の高い割合で回文を生成できた。これは、本研究の提案手法のほうが通意条件を満たした回文を生成できることを意味する。

単語列変形処理を加えたことによって、bi-gram に含まれる文節ペアの個数が増加し、回文の探索空間が著しく広がると

表 3: 変形処理を加えたモデルが生成した文章の例

モデル	生成回文	シード文節	モデル	生成回文	シード文節
B+省略	うその 答え 得た この 僧	答え	B+倒置	理解の ある 今 いる あの 怒り	今
	遠く 行く 音	行く		捜した 私が さ	私が
	汁 味 ある 死	味		実る 秋 ある 飲み	秋
B+置換	ないなら 今 いらないな	今	B+全部	確か 一度 ほど 血 生かした	ほど
	害 あり 知り合いが	知り合いが		リスク 残る この 薬	残る
	得た ことは と 答え	ことは		値 張る 味 ある 羽	味

表 2: モデル B が生成した回文の一覧

生成回文	シード文節
行けとの こと この 時計	こと
さ あの 日の 朝	日の
日のことこの日	こと
試合が なく 長い 足	なく

同時に、回文の生成にかかる時間も増加した。しかしながら、コーパスの文を変形し、回文の探索空間を広げても、通意条件を満たす回文の割合（Y/X）はさほど減少していない。例えば、「B」と「B+全部」を比較すると、生成される回文候補の数は 43.2 倍に増えたが、通意条件を満たす確率は 0.39 倍の減少にとどまっている。これは，提案手法による文の変形が、通意条件を満たしながら回文の特徴をとらえた変形になっていることを示している。

表 2 に、モデル B が生成した回文を示し、表 3 に変形処理を加えた他のモデルが生成した回文の例を示した。以下、文節に施した各変形処理が、文節 bi-gram を通して生成回文に与えた影響について述べる。

4.1 省略処理:モデル B+省略

表 3 より、モデル B+省略がモデル B+全部を除けば最も多くの回文を生成したとわかる。これは、「を」「が」「は」を省略することで、音の条件を満たすことが容易になったことを示している。例えば、モデル B+省略は「うその 答え 得た この 僧」を生成した。これは、文節「僧が」や「僧は」から助詞を取り除いたことで生成可能になった例である。

また、モデル B+省略は、モデル B+全部を除いた中では最も多くの音の条件を満たす文を生成した。その上、候補に対する回文の割合の減少は小さく抑えられている。つまり、助詞の省略処理は、通意条件を満たしたまま音の条件を達成するのに非常に効果的である。

その一方で、モデル B+省略は、音の条件を満たすが意味を成さない文として「汁 味 ある 死」を生成した。これは助詞の省略処理を加えた文節ペアばかりが bi-gram から選ばれていることが原因である。

4.2 倒置処理:モデル B+倒置

表 3 より、移動処理によって、「理解の ある 今 いる あの 怒り」が生成できるようになった。素 bi-gram にあった「今 ある」のペアは、係り受け構造から「ある 今」へと変形されたことにより生成された例である。結果、シード文節「今」に対して、「ある」を左側に結合することが可能となり、回文が生成された。

同様に、「捜した 私が さ」も、文節ペア「私が 捜した」を「捜した 私が」に変形することで生成が可能になった。

4.3 置換処理:モデル B+置換

表 3 より、活用形処理を加えたことで「害 あり 知り合いが」が生成できた。これは、素 bi-gram に含まれていた「害 ある」の文節ペアを、「ある」の活用形「あり」で置換したことによるものと考えられる。結果、bi-gram に「害 あり」が新たに追加され、文節「あり」の左側に「害」が連結可能となった。

同様の例としては表 3 より回文「ないなら 今 いらないな」が挙げられる。ここでも、基本形「いる」を「いらないな」に置換することで、回文が生成された。

5. まとめと今後の課題

本研究では、文節の bi-gram を用いた折り返し固定法による、文節間の依存関係を考慮した回文の自動生成手法を提案した。単純な文章データから bi-gram を構築するだけでは、省略や倒置などの回文独自の表現手法に対応できないため、単語列の変形規則を適用することで、bi-gram を拡張した。

その結果、生成した文章のうち、音の条件を満たした文章に対する回文の割合を、既存研究の 1,000 倍以上に引き上げることに成功した。また、省略、倒置と置換の変形処理を加えた各 bi-gram モデルからは、それぞれの表現を含む回文を生成することができた。つまり、変形規則を bi-gram に適用することで、回文特有の文体を考慮した回文を生成できるとわかった。また、コーパスの文の変形処理を適用したことにより、通意条件を満たす割合をさほど低下させることなく、回文の探索空間を広げることができた。

現状は、文節 bi-gram の連接情報のみを用いており、コーパス中の出現確率などは考慮していない。今後は、文節 bi-gram の連接情報を、より一般的な文節 n-gram 言語モデルに拡張し、bi-gram よりも長い範囲の特徴や、コーパス中の出現確率を考慮した回文探索アルゴリズムを考えていきたい。

参考文献

[1] 鈴木啓輔, 佐藤理史, 駒谷和範. 文頭固定法による効率的な回文生成. 言語処理学会 第 17 回年次大会 発表論文集, pp. 826–3829, 2011.

[2] 鈴木啓輔, 佐藤理史. 文節結合による回文の自動生成. In *The 24th Annual Conference of the Japanese Society for Artificial Intelligence, 2010*, 2010.