事象の連想と共起性に基づいたマルコフ連鎖による文生成

Generating Sentences with Markov Chain based on Association and Co-occurrence

秋山 陸，寺岡 丈博
Riku AKIYAMA, Takehiro TERAOKA
拓殖大学工学部情報工学科
Department of Computer Science, Faculty of Engineering, Takushoku University
Email: r88406@st.takushoku-u.ac.jp

あらまし：本研究では，用意した文章から計算された単語間の推移頻度に，連想概念辞書を用いて重み付けを行う．このデータを用いて事象の連想と共起性の両方の側面から，人間が書いたような文章を生成することを目的とする．評価実験では，マルコフ連鎖によって生成された文章と，提案手法によって生成された文章を比較した．その結果，提案手法を用いることでより自然な文章が生成されることを確認した．
キーワード：自然言語処理，文章生成，マルコフ連鎖，連想概念辞書

1. 背景と目的

近年，自然言語処理や人工知能の分野において，文章を自動生成する研究が関心を集めている．文章の自動生成は自然言語処理の分野でも難易度が高いタスクであり，日々研究が進められている．

本研究は，既存の文章コーパスから計算されたマルコフ連鎖のモデルに，連想概念辞書[1,2]のデータで重み付けをすることで，事象の連想と共起性の両方の側面から，人間が書いたような自然な文章を生成することを目的とする．

2. 関連研究

高木らは，物語全体のプロットを用意し，指定の箇所に単語や単文を導入することで物語小説の自動生成を試みている[3]．しかし，文章の生成には人手により作成されたプロットを必要とするため，自動で話全体を組み立てることに関してはできていないといえる．

吉田らは，まず入力文中の単語に対して，文法を崩さずに置換できるボケ単語を Google N-gram から探して置換することで，ボケの文章を生成する．その後，置換前の単語をツッコミのテンプレートに当てはめることで，ボケとツッコミのセットになった対話文の生成を試みている[4]．しかし，ボケは入力文をテンプレートとして，ツッコミは用意されたテンプレートから生成されるため，文章の大部分を自動で生成できてはいないといえる．

これらの点を踏まえて，本研究では文章として意味が自然に読むことができ且つ生成する際になるべく人手を介さずに文章を生成できる手法を提案する．

3. 提案手法

3.1 手法の概要

図 1 は手法の概要を表している．まずはじめに，新聞コーパスからタグや見出し，改行やスペースを取り除いて整形を行い，本文のみを抽出する．抽出された小説の本文を，Janome を用いて分かち書きを行い，マルコフ連鎖のための推移確率テーブルを作成する．このマルコフ連鎖のテーブルに連想概念辞書の連想距離を用いて重み付けを行い，本研究で目的とする人間の連想を取り込んだマルコフ連鎖のテーブルデータとする．

このテーブルデータを用いて文章を生成することで，より人間的で意味が通りやすい文章になると考えられる．

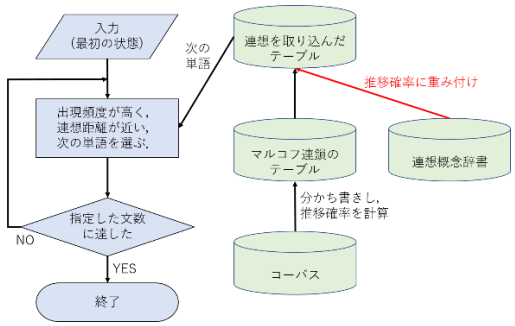

図 1 手法の流れ

3.2 連想概念辞書

連想概念辞書とは，動詞（刺激語）から連想された語（連想語）が収録されている辞書である．辞書には，表 1 のようなデータが 30 万行ほど記録されている．

表 1 連想概念辞書

刺激語	連想語	連想距離	連想スコア
作る	必死に	3.79	0.26
作る	しっかりと	6.28	0.14
作る	一気に	8	0.12
作る	手を抜いて	9	0.11

表 2 出力結果

従来手法	提案手法
力を生かし、常に「アフリカ系の人に申し訳ない。	力を生かし、ＦＷ戦で攻め続けている」と語った。
力を生かし､常に２､後藤大津２１９９０･６％前後」。	力を発揮できた。
相手を崩せず完敗を喫し、安倍首相は、平成最後の年の瀬に 生まれた赤ちゃん。	相手を崩せず完敗を喫し､安倍首相は退陣に追い込まれた。

本研究では「刺激語」「連想語」「連想距離」の項目を用いる．連想距離とは，刺激語と連想語の 2 単語間の連想の強さを表す概念で，この数値が小さいほど強く連想されやすい．提案手法では，連想が強まるほど数値が大きくなるように，連想距離の逆数を「連想スコア」という新たな項目として辞書に付与し，これを重み付けに用いる．

3.3 連想概念辞書による重み付け

新聞コーパス 10 万字分からマルコフ連鎖のテーブルを生成したところ, 表 3 のような trigram の組み合わせが 46378 組生成された．そのなかで，現在の状態である左の 2 単語と，次の状態（次の単語の候補）との組み合わせが，連想概念辞書の刺激語と連想語の組み合わせとして，426 組存在した．

これらの組み合わせには，推移確率に連想距離の逆数を足すことで，連想が強いほど次の状態として選択される確率が上がるように重み付けを行った．（例：「刺激語：作る」と「連想語：一気に」の連想距離の逆数は 0.12 のため，「作っ，て」の次に「一気に」に推移する確率の 0.33 と足すと 0.45 となる．）

表 3 推移確率と連想スコア

現在の状態		次の状態	推移確率	推移確率+連想スコア
作っ	て	もらっ	0.33	0.33
作っ	て	い	0.33	0.33
作っ	て	一気に	0.33	0.45
て	もらっ	た	0.6	0.6
て	もらっ	て	0.4	0.4

4. 結果と評価

4.1 結果

従来手法（連想を取り入れていない 2 階マルコフ連鎖）と本研究の提案手法での文章生成結果を比較すると，表 2 のようになった．

4.2 評価実験

従来手法と本研究の提案手法で生成された文章を比較するため，最初の状態（文頭の 2 単語）を固定した上で，各手法で生成した単文で「どちらのほうが」かどうかを，アンケートを用いて評価する．

評価実験の結果は表 4 のようになった．「力，を」から始まる文章群については，「提案手法の方が，自然な文章である」という回答が 62.5%，「相手，を」から始まる文章群については，どちらも同じ 50%という結果になった．提案手法の評価が従来手法の評価を下回ることはなく, 提案手法を用いることで「人間の発想に近い自然な文章」の自動生成に一定の成果があるという結果が得られた．

表 4 実験結果

	従来手法	提案手法
最初の状態（力，を）	37.5%	62.5%
最初の状態（相手，を）	50%	50%

5. 考察

出力結果の変化は見られたものの，文章の自然さにおいて大幅な改善とまでは言えない結果になった．

考えられる理由として，マルコフ連鎖のテーブルに対して連想概念辞書を適用できる部分が少なかったことが挙げられる．46378 項目中 426 項目に連想概念辞書から，重みを付与することができたが、これはテーブル全体の 1%程度しか適用できなかったことになるため，大幅な改善につながらなかったと考えられる．

また，日本語の文章は接続詞や助詞や句読点などが多いため，前後 3 つの単語について連想関係を考慮するだけでは範囲が狭く，もっと文全体に連想関係を適用することができれば改善の余地があると思われる．

連想概念辞書の性質上，文章が堅く固有名詞が多い新聞よりも，SNS やブログのような砕けた文章をベースにした方が，より精度が改善される可能性がある．

謝辞

本研究は JSPS 科研費 JP18K12434 の助成を受けたものです．

参考文献

(1) 寺岡丈博，東中竜一郎，岡本潤，石崎俊：“単語間連想関係を用いた換喩表現の自動検出”，人工知能学会論文誌，Vol. 28, No. 3, pp. 335-346(2013).

(2) T. Teraoka: "Analysis of Associative Information for Second Language Learning of Japanese", In *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference (APCLC)*, pp. 434-439(2018).

(3) 高木大生，佐藤理史，松崎拓也：“プロットと背景知識を用いた短編小説の自動生成”，情報処理学会第 77 回全国大会講演論文集, pp. 171–172 (2015).

(4) 吉田裕介，萩原将文：“漫才形式の対話文自動生成システム”，日本感性工学会論文誌，Vol.11，No.2，pp. 265–272 (2012).