文書生成のための文の並べ替え

大田 浩志 , 山本 和英
長岡技術科学大学 電気系
E-mail:{ota, ykaz}@nlp.nagaokaut.ac.jp

1　はじめに

自然な文書を生成することは自然言語処理の大きな課題である。1 文で表現できる情報は限定的であり、多くの場合文書としての出力が求められる。しかしながら、文書生成の研究は多いとは言えない。その原因として文脈を扱う問題の難しさがある。我々は統計情報により文脈をとらえることを試みる。本研究では文書生成を目的として、複数の文を 1 つの文書として尤もらしくなるように並べることを考える。

文を並べ替えて 1 つの文書を生成する試みには複数文書要約を目指したものがある。これは複数の文書から文を抽出して 1 つ文書を生成する場合、文の並び順が文書の読みやすさや自然さに影響を与えるためである [1]。これらの研究では要約対象となる文書 (原文書) の情報を用いて文の並び替えを行っているものが多い。具体的には原文の書かれた順序、原文における位置などを用いている [2]。しかし我々は要約ではなく文書生成に興味がある。原文書の存在しない文、例えば文生成システムにより生成された文を対象に並べ替えが出来れば文書を生成することができる。そこで本稿では原文書の情報を用いないで文の並べ替えを行うことを目指す。具体的には既存の文書の文を入力として、元通りにできるかを実験した。統計情報を用いて文の並べ替えを行った結果を報告する。また、並べ替えの対象の違いによる異なりを考察する。

本研究では統計情報を用いることで 2 文間の順序関係を推定し、複数の文を並べ替える。このような文同士の関係を推定する研究はまだ少ない。

鎌田ら [3] はストーリーの自動生成を行なっている。その際、把握しやすさを考慮するために、シソーラスを用いて文の並べ替えを行うことで意味のまとまりとしての読みやすさを得ようとしている。また山本ら [4] は隣接文の結束性を自動判別する手法を提案している。これは談話の整合性を考慮した要約を行うことを目的としている。結束性の判定に有効だと考える素性を用いて、SVM を用いて学習し判別器を構築している。素性として接続詞や単語間類似度の情報を用いている。Lapata[5] は本研究と同様に統計的手法により文の並べ替えを行っている。また Bollegala et al.[6] は単純な文の並べ替え手法を組み合わせることで個々の精度より良い結果が得られることを報告している。

2　研究対象

本研究では、性質の異なる 2 つの文書において文の並べ替えを試みる。2 つの文書とは新聞記事及びレビュー記事である。具体的には日本経済新聞 (2) 及び Amazon レビュー (3) を用いる。新聞記事の場合、書き手は訓練された限られた人でありリード文を含むなど形式の定まった文書であるといえる。それに対して、レビュー記事の形式は比較的自由である。近年ウェブ上では個人が自由にレビューを記述している。言語資源としての多様性は新聞記事に勝るだろう。

奥村は意見の含むものを対象にした要約の必要性を述べている [7]。また難波は主観表現を含む複数文書要約が必要になっていると述べている [8]。レビュー記事は主観表現を含む文書である。

そこで我々はレビュー記事の並べ替えは自由度の高い文書生成に繋がると考え、これら 2 つの文書を対象に提案する文の並べ替え手法の有効性を確認する。

3　人手による文の並べ替え

新聞記事及びレビュー記事を対象に人手による文の並べ替え実験を行った。Barzilay et al.[1] は新聞記事を対象に並べ替えを行い、文の並び順が文書の読みやすさに影響することを報告している。我々はレビューにおいても同様のことがいえるかを確認する。本節ではまず、レビュー記事においても尤もらしい文の並び順があることを確認する。また新聞記事を対象とした場合の実験結果と比較することで両記事の性質の違いを確認する。

実験は被験者 1 名により実施した。既存の記事 (原文書) の順序を無作為に入れ換え、被験者に提示した。被験者は提示された数文を 1 つの文書として読みやすくなるように文を並べ替える。また被験者は並べ替えを行った記事に対して自己評価を行う。評価値は以下の 3 つである。

評価 (1) この順序以外では読みにくい
評価 (2) 他の順序でも読めるがこの順序が尤もらしい
評価 (3) 順序を持たない文を含む

実験対象として、5 文で構成された新聞記事及びレビュー記事を無作為に 10 件ずつ用意した。ただし、いずれの記事も指示詞及び接続詞を含まないものの中から選んだ。これは、これらの単語を含む文と含まない文とではその並べ替えやすさが異なる為である。表 1 に評価値毎の記事数を示す。

表 1: 各評価値を得た記事数

対象記事	評価 (1)	評価 (2)	評価 (3)
新聞	4	6	0
レビュー	1	5	4

表 1 よりレビュー記事は新聞記事と比較して、評価 (3) が多いことがわかる。被験者による文の並び順と原文書の文の並び順との相関を確認する。評価指標はケンドールの順位相関係数 τ(式 (1)) を用いる。Lapata[9] はケンドール相関係数と人間にとって尤もらしい文の並び、および読むのにかかる時間との相関があると報告している。

$$\tau = \frac{4I}{N(N-1)} \tag{1}$$

ただし、N は文の数、I は原文における文の順序と並べ替えによる文の順序が反転している文対の数である。表 2 に各記事毎の相関値を示す。

表 2 よりレビュー記事は新聞記事と比較して原文書の文の並び順との相関が低いことがわかる。表 1 及び表 2 より新聞記事と比較して、レビュー記事は文の並びが自由な文書であると言える。

次に文数の異なりにより並べ替えの困難さが変化するかを確認する。並べ替えの対象は Amazon の書籍カテゴリの

表 2: 原文書の順序と人手実験による順序の相関 (対象記事の違い)

対象記事	τ	最小	最大
新聞	0.66	0.20	1.00
レビュー	0.42	-0.60	1.00

レビューとし、3 行/5 行/7 行で構成される各 10 記事を被験者に提示した。表 4 に評価結果をまとめる。

表 3: 原文書の順序と人手実験による順序の相関 (文数の違い)

文数	τ	最小	最大
3	0.33	-1.0	1.0
5	0.66	-0.6	1.0
7	0.59	0.14	1.0

文数が多い方が最小値が大きいことがわかる。短い文数の場合、人によって文の書き順が大きく異なる。この結果を踏まえて自動で文の並べ替えをすることを考える。

4 統計情報を用いた文の並べ替え

統計情報を用いた文の並べ替えの先行研究として Lapata[5] の提案する確率モデルがある。本節では、まず Lapata の提案するモデルについて述べ、我々の提案手法との一致点及び相違点を述べる。さらに 4.2 節,4.3 節では提案する文の連接しやすさ/しにくさの尺度について述べる。

4.1 Lapata の確率モデル

Lapata は確率モデルによる文の並べ替えを提案している。確率モデルは既存の文書を学習データとし、1 文を構成する素性 (例えば、単語) が連続した文に出現する確率を用いている。

n 文で構成される文書 T を対象に文 S を並べ替える。S_i は i 番目の文を表す。

$$T \in (S_1, S_2, ..., S_n) \tag{2}$$

2 文間の連接確率 $P(S_i|S_{i-1})$ を用いて文書全体での文の連接確率 $P(T)$ が最大となるように文を並べ替える。

$$P(T) = \prod_{i=1}^{n} P(S_i|S_{i-1}) \tag{3}$$

$$P(S_i|S_{i-1}) = \prod_{(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle}) \in S_i \times S_{i-1}} P(a_{\langle i,j\rangle}|a_{\langle i-1,k\rangle}) \tag{4}$$

ここで $a_{\langle i,n\rangle}$ は文 S_i を構成する素性である。

$$S_i \in (a_{\langle i,1\rangle}, a_{\langle i,2\rangle} ... a_{\langle i,n\rangle}) \tag{5}$$

素性 $a\langle i,j\rangle$ と $a\langle i-1,k\rangle$ が連続する 2 文に出現する確率は次式で与えられる。

$$P(a_{\langle i,j\rangle}|a_{\langle i-1,k\rangle}) = \frac{f(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle})}{\sum_{a_{\langle i,j\rangle}} f(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle})} \tag{6}$$

$f(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle})$ は素性 $a\langle i,j\rangle$ が $a\langle i-1,k\rangle$ の次の行に出現する頻度である。

我々の手法においても式 (6) と同様に 2 単語間に連接確率を付与する。Lapata は単語間の連接確率を文間の連接確率に拡張する際、その総積により算出している (式 (4))。しかし我々は 2 単語間の連接確率の総和により文間の連接のしやすさを表すスコアを算出する。これを $Connect_\beta(S_{i-1}, S_i)$ とし、4.2 節で説明する。総積により算出した場合、極端に連接確率の低い単語対により文全体の連接確率が下がることになる。これは学習データに含まない語が出現した場合に影響が大きいと考えた。

単語の連接確率の総和により文間の連接を算出すると、文が長くなるほどスコアが高くなる。これを抑止するために、2 単語間にはもうひとつ統計情報を付与する。$Connect_\beta(S_{i-1}, S_i)$ が連接しやすさを意図しているのに対して、連接しにくさを表す $Un\_Connect(S_{i-1}, S_i)$(式 (7)) を定義する。4.2 節で説明する。ただし α は重みである。

$$\begin{aligned} &Connect(S_{i-1}, S_i) = \\ &Connect_\beta(S_{i-1}, S_i) - \alpha Un\_Connect(S_{i-1}, S_i) \end{aligned} \tag{7}$$

4.2 2 文の連接しやすさの尺度 $Connect_\beta$

2 文間に含まれる単語の 2 文間における連接確率を用いて算出する。S_{i-1} の次の文が S_i のときの連接しやすさ $Connect_\beta(S_{i-1}, S_i)$ は次式で表す。

$$\begin{aligned} &Connect_\beta(S_{i-1}, S_i) = \\ &\sum_{(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle}) \in S_i \times S_{i-1}} P(a_{\langle i,j\rangle}|a_{\langle i-1,k\rangle}) \end{aligned} \tag{8}$$

連接確率に基づくスコアなので文の並びによりスコアは異なる。

$$Connect_\beta(S_{i-1}, S_i) \neq Connect_\beta(S_i, S_{i-1}) \tag{9}$$

4.3 2 文の連接しにくさの尺度 $Un\_Connect$

2 文間の連接しにくさを算出する手法を提案する。連続する 2 文における 2 単語の共起情報と、1 文書における 2 単語の共起情報を用いる。1 文書内で共起しやすいが、連続する 2 文においては共起しにくい単語対は連接しにくさを表す。文書 d における共起しやすさを表す指標として相互情報量 (式 12) を用いる。ただし N_d は統計情報を得るために用いた文書 d の総数である。1 文書内における単語の共起スコア Co_oc_{txt} と 2 文内共起スコア Co_oc_{2sent} の差を 2 単語の連接しにくさ $Un\_Connect_w$ とする。2 文間の連接しにくさ $Un\_Connect(S_{i-1}, S_i)$ を式 (10) に示す。

$$\begin{aligned} &Un\_Connect(S_{i-1}, S_i) = \\ &\sum_{(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle}) \in S_i \times S_{i-1}} Un\_Connect_w(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle}) \end{aligned} \tag{10}$$

$$\begin{aligned} &Un\_Connect_w(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle}) = \\ &Co\_oc_{txt}(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle}) - Co\_oc_{2sent}(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle}) \end{aligned} \tag{11}$$

$$Co\_oc_d(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle}) = -log_2 \frac{f(a_{\langle i,j\rangle}) f(a_{\langle i-1,k\rangle})}{f(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle}) N_d} \tag{12}$$

5 評価実験

提案手法の有効性を確認するために実験を行う。本稿では並べ替えに用いる素性は名詞-一般及び名詞-固有名詞 (1) とした。ただし、名詞-固有名詞はタグにより同一化した。

連接確率及び共起確率の算出に用いる学習データを 2 種類用意する。学習データは日本経済新聞 1996-2000 年及び Amazon レビューである。並べ替えの対象とする実験データとして、3 行/5 行/7 行でそれぞれ構成された新聞記事、レビュー記事を各 100 記事用意した。新聞記事の並べ替えに用いる統計情報は新聞記事を学習データとした。同様に並べ替え対象がレビュー記事の場合はレビュー記事を学習データとした。提案手法、Lapata の手法及びランダムで並べ替えた結果を表 4 に示す。

表 4: 原文書の文の並びと各手法による文の並びの相関値 (τ)

対象記事	手法	3 文	5 文	7 文
	ランダム	0.02	-0.01	0.01
新聞	提案手法	0.01	0.02	0.01
	Lapata	0.22	0.15	0.07
レビュー	提案手法	0.07	0.15	0.12
	Lapata	0.17	0.13	0.11

学習データの違いによる並べ替え結果を比較するための実験を行なう。5 文で構成される新聞記事を対象に次の 3 種類の学習データを用いて並べ替えた。3 種類とは、並べ替え対象記事を含む新聞記事 (Closed)・並べ替え対象記事を除いた新聞記事 (Open)・レビュー記事 (Open_d) である。レビュー記事に対しても同様に 3 種類の学習データを用意した。並べ替え対象記事を含むレビュー記事 (Closed)・並べ替え対象記事を除いたレビュー記事 (Open)・新聞記事 (Open_d) である。実験結果を表 5 に示す。

表 5: 原文書の文の並びと各学習データによる文の並びの相関値 (τ)

対象記事	手法	Closed	Open	Open_d
新聞	提案手法	0.02	-0.03	-0.02
レビュー	提案手法	0.15	0.10	-0.05
	Lapata	0.13	0.10	-0.06

6 考察

6.1 並べ替え対象の文数の異なりによる結果の比較

異なる文数の文書を並べ替えた結果を比較する。表 4 より既存手法では文数が多い文書程、原文書との文の並びの相関が低くなっていることがわかる。しかしながら提案手法では 3 文で構成される文を並べ替えた場合の相関値が最も低くなった。この原因を $Un\_Connect$ の有効性に注目して考える。式 (13) で表す $Un\_Connect(S_i, S_{i\pm a})$ を算出する。

$$Un\_Connect(S_i, S_{i\pm a}) = \frac{1}{2}(Un\_Connect(S_i, S_{i-a}) + Un\_Connect(S_i, S_{i-a})) \quad (13)$$

a は原文書中の 2 文間の距離を表す。$Un\_Connect$ は連接しにくさを意図したスコアである。つまり、原文書における文間の距離に比例してスコアが高くなるのが理想的である。5 文で構成された記事を対象に $Un\_Connect(S_i, S_{i\pm a})$ を算出した。結果を表 6 に示す。

表 6: 文同士の距離と $Un\_Connect$

	$Un\_Connect(S_i, S_{i\pm a})$			
対象記事	$a=1$	$a=2$	$a=3$	$a=4$
新聞	1.31	1.40	1.37	1.45
レビュー	2.05	2.08	2.15	2.15

表 6 より、新聞記事/レビュー記事の双方において $a=1$ のとき最も低い値をとっていることがわかる。原文書における文間の距離が近い 2 文程スコアが低いことから、$Un\_Connect$ は意図したスコア付けが出来ている。しかしながら、レビュー記事において $a=3$ と $a=4$ には差がみられない。このことから原文書における文間の距離に距離に比例してスコアが高くなっているとはいえない。

表 4 では 3 文を並べ替えた相関値は 5 文/7 文のものと比べて低い。3 文の記事の場合、最も離れた文同士でも 2 文 ($a=2$) しか離れていない。それに対して 5 文を対象とする場合は最大で 4 文 ($a=4$) 離れている。$a=2$ と $a=4$ を比較すると $a=4$ の $Un\_Connect$ スコアが高くとなっている。これより文対 S_1, S_5 は S_1, S_3 と比較して連接しにくいことがわかる。原文書において最も離れている 2 文が連接した場合、相関値は大きく下がるだろう。3 文の記事を対象とした場合、5 文の記事を対象とする場合に比べて最も離れている 2 文が連接しやすくなると考える。その結果 3 文と比較して 5 文の並べ替えの結果が良くなったと考える。

また $a=3$ と $a=4$ を比較すると差がないことがわかる。これより a が大きくなるほどスコアが大きくなるとはいえない。そのため $Un\_Connect$ は 7 文の文書に対して 5 文の文書程効果的でなかったと考える。

6.2 新聞記事のレビュー記事の比較

表 4 に示す実験結果より、提案手法を新聞記事に適用した場合並べ替えにはほとんど効果がないことがわかる。本手法により並べ替えがしにくい文があるのではないかと考えた。

本手法により並べ替えが正しく行なわれるためには、$Connect(S_1, S_2)$ が $Connect(S_1, S_3), ..., Connect(S_1, S_5)$ に比べて値が大きくなっていなければならない。ここで S_1,S_2 のように原文書における正しい並びの 2 文に対して付与されるスコアを $Connect(S_i, S_{i+1})$ と表す。それに対して、S_1,S_3 のように原文書では連接しない 2 文のスコア $Connect(wrong)$ を式 (14) で表すことにする。

$$Connect_i(wrong) = \frac{1}{3}\sum_{j=1}^{5} Connect(S_i, S_j) \quad \text{但し } j \neq i, i+1 \quad (14)$$

ここでは 5 文で構成する新聞記事及びレビュー記事において、$Connect(S_i, S_{i+1}) > Connect(wrong)$ となっているかを確認する。結果を表 7 に示す。

表 7 より新聞記事を並べ替えの対象としたときの $Connect(S_1, S_2)$ が $Connect(wrong)$ に比べて小さいことがわかる。つまり 1 文目及び 2 文目の連接の強さが低いということを示している。これには新聞の 1 文目に書かれるリード文による影響があると考える。リード文は記事の要

表 7: 何文目と何文目の連接が誤っているか

	$Connect(S_i, S_{i+1}) - Connect(wrong)$			
対象記事	$i=1$	$i=2$	$i=3$	$i=4$
新聞	-0.06	0.13	0.094	0.13
レビュー	0.014	0.019	0.028	0.039

約であるため記事全体を構成する単語を含んでいる。つまりリード文に含まれる語は 2 文目以降の文とも繋がりが強い場合があり、並べ替えしと考える。

6.3 既存手法と提案手法の比較

表 4 の新聞記事を対象とした結果をみると、Lapata の手法による結果の方が良いことがわかる。なぜこの差が生じたかを考える。我々の提案手法は Lapata の提案する手法と以下の 2 点で異なる。

- 2 文に含まれる単語間の連接確率をスコアとみなし総和をとることで 2 文の連接しやすさとしている
- *Un_Connect* を取り入れている

表 6 より *Un_Connect* には大きな問題はないことは 6.1 節で述べた。そこで、連接確率スコアの総和としたところに問題があったと考える。単語の連接確率の総和をとるということは連接スコアの高い単語対の影響が大きくなる。

新聞記事を対象とした場合に並べ替えが出来ていない。この結果から連接スコアの高い単語対のみでは文間の連接を表せていないということがわかる。原因のひとつとして新聞記事においては固有名詞が多く使われていることがあげられる。固有名詞は頻出単語な上、同一化しているため頻度が高い。そのため、2 文間に固有名詞を含む場合その *Connect* スコアは高くなると考える。同一化した固有名詞により文の順序は定められないと考えられ、本手法では並べ替えが出来なかったと考える。

6.4 学習データと並べ替え対象の関係

表 5 に示す結果について述べる。レビュー記事で学習しレビュー記事を並べ替えたクローズドテストのとき最も良い結果 (相関値 0.15) が得られた。学習データをレビュー記事としたときレビュー記事を対象としたオープンテストにおいても相関値 0.1 となり、並べ替えに効果があることを示している。

しかしながら、レビュー記事の並べ替えに用いる学習データを新聞記事としたとき、並べ替えの効果はみられなかった。同様に新聞記事の並べ替えに用いる学習データをレビュー記事としたとについても効果はみられない。このことから統計情報は日本語文書の文脈をとらえているとは言えず、学習データとした文書それぞれの特徴をとらえているだけであると考える。

さらに学習データとしてレビュー記事及び新聞記事をまとめて用いてレビュー記事の並べ替えを行なった。その結果は相関値は 0.004 となり、並べ替えが出きているとは言えない。並べ替え対象と同じ性質を持ったもののみで学習する必要があると考える。

7 まとめ

統計的手法を用いて複数の文を 1 つの文書として尤もらしくなるように並べ替える手法を提案した。並べ替え実験の結果レビュー記事を対象として並べ替えたとき、原文書の並びとの相関値は 0.15 であった。これは十分に並べ替えができているとはいえない。しかし、既存手法にはない単語共起に基づく文間の連接しにくさを表すスコア *Un_Connect* を提案しその有効性を確認した。

Bollegala et al.[6] は単純な手法を組み合わせることで個々の精度よりも良い結果が得られることを報告している。これは文を並べ替えるのには様々な要因が影響していてるからだと考える。単一のモデルで文の並べ替えを実現するのは困難である。提案手法では既存手法と比較して異なる結果が得られている。よって本手法の結果も素性のひとつとして有効だと考える。

また新聞記事とレビュー記事の人手での並べ替え実験を行なうことで、レビュー記事は新聞記事と比較して文の順序の自由度が高いことがわかった。そして並べ替え対象記事と統計情報の学習データは同じ性質の文書にする必要があることがわかった。

使用した言語資源及びツール

(1) IPA 品詞体系日本語辞書 ‘‘IPADIC’’,Ver.2.7.0, 奈良先端科学技術大学院大学 松本研究室, http://chasen.naist.jp/stable/ipadic/

(2) 日本経済新聞全記事データベース 1996-2000 年度版, 日本経済新聞社

(3) Amazon.co.jp,http://www.amazon.co.jp/

参考文献

[1] Regina Barzilay, Noemie Elhadad, and Kathleen R. Mckeown. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, Vol. 17, p. 2002, 2002.

[2] Kathleen R. Mckeown, Judith L. Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. Towards multidocument summarization by reformulation: Progress and prospects. In *In Proceedings of AAAI-99*, pp. 453–460, 1999.

[3] 鎌田健一. 雑多なテキスト集合からのストーリー生成. 言語処理学会第 8 回年次大会論文集, pp. 363–366, 2002.

[4] 山本悠二, 増山繁, 酒井浩之. 小説自動要約のための隣接文間の結束性判定手法. 言語処理学会第 12 回年次大会論文集, pp. 1083–1086, 2006.

[5] Mirella Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *In Proc. of the Annual Meeting of the Association for Computational Linguistics*, pp. 545–552, 2003.

[6] Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. A machine learning approach to sentence ordering for multidocument summarization and its evaluation. IJCNLP.

[7] 奥村学. Tsc4： 意見要約コーパスとそれを用いたワークショップ. 言語処理学会第 11 回年次大会 (NLP2005), 2005.

[8] 難波英嗣. 情報抽出を利用した複数文書要約. 知能と情報： 日本知能情報ファジィ学会誌： journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol. 18, No. 5, pp. 682–688, 20061015.

[9] Mirella Lapata. Automatic evaluation of information ordering: Kendall’s tau. *Computational Linguistics*, Vol. 32, No. 4, pp. 471–484, 2006.