# 教師あり機械学習を用いた段落の順序推定

伊藤 聡史 [*1]　村田 真樹 [*2]　徳久 雅人 [*2]　馬 青 [*3]
[*1] 鳥取大学 工学部 知能情報工学科
[*2] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
[*3] 龍谷大学 理工学部 数理情報学科
[*1,*2]{s092007,murata,tokuhisa}@ike.tottori-u.ac.jp
[*3] qma@math.ryukoku.ac.jp

## 1　はじめに

われわれが文章作成を行う際，読者が読みづらい文章を作成することがある．読みづらい文章には，意味の分からない専門用語を用いることや，狭い文章中に複数の話題が存在すること，冗長な文章を用いること，文章の順番が良くないことなど，様々な原因が存在する．本稿では，そのうちの文章の順番の問題を取り上げる．

文章の順番が良くないために読みづらい文になっている場合は，文章を適切な順序に並べ替える必要がある．文章を適切な順序に並べ替えるために，本稿では教師あり機械学習を利用する．教師あり機械学習には性能が高いと広く認識されているサポートベクトルマシン (*SVM*) を用いる．

機械学習を用いた文章の順序の推定として，内元ら [1] や林ら [2] の研究がある．内元らは単語の順序，林らは文の順序を扱った．このため，本稿では段落の順序の推定を行う．

本稿では，2 段落ごとで元の順番 (正順)・その逆の順番 (逆順) の 2 通りの場合についての問題を作成し，機械学習を用いることにより，どちらの順序が正しいかを判定する．さらに，段落の順序推定に役立つ素性の分析も行う．

本稿の特徴を以下に整理する．

- 段落の順序の推定に，教師あり機械学習を用いているという特徴がある．
- 教師あり機械学習を用いることにより，新たに素性を低コストで，かつ大量に組み込むことができる．性能向上に有用な素性が見つけることができる可能性がある．
- 記事の最初の 2 段落における順序推定では，教師あり機械学習を用いる提案手法で 0.85 という高い正解率を得た．この正解率は人手の正解率と同等であった．
- 記事内の連接する 2 段落における順序推定では，提案手法で 0.60 の正解率であった．前方の文章との名詞の一致数が大きい段落を前方とするベースライン手法より高い正解率であった．
- 文の順序推定と段落の順序推定の比較を行い，違いを明らかにした (5.5 節)．

## 2　関連研究

内元らは文生成のために，最大エントロピー法を用いて文節の係り受け情報をもとに単語の順序を推定する研究を行った [1]．正しい語順をコーパス内での語順とすることにより，語順に関わる学習データをコーパスから自動的に構築でき，人手での学習データの作成を不要としている．

林らは新聞記事から文の順序推定のために，多数の素性を用いた教師あり機械学習に基づく研究を行った [2]．新聞記事から 2 文一組で抜き出し，その 2 文から元の順の文（正例）と逆順の文（負例）を作成し教師あり機械学習を用いてその 2 文が正例か負例かを判定して文の順序を推定するというものである．機械学習で用いるデータは，内元らの研究を参考にしてコーパスから自動で構築できるようにした．実験では，段落内最初の 2 文のみを用いる場合と，段落内全ての連接した 2 文を用いる場合と，段落内全てから 2 文を用いる場合の 3 種類における順序推定を行った．さらに，Lapata[3] の手法に基づく確率手法と比較をした．比較実験により林らの手法の方が優れた性能であったと報告された．

これらの研究では，文節/文の順序推定を扱った．これらに対して，本稿は段落の順序推定を扱うという違いがある．

Lapata は既存する文章を学習データとし，文に含まれる素性が連続した文に出現する確率を求めている．その値の総積により 1 文目に対し 2 文目が配置される確率を算出し，その確率に基づき文の順序を決定する研究を行った [3]．文の順序推定には，2 文間の動詞の順序性や名詞の同一性，文構造などを用いている．

横野らは，テキスト内に一貫性の良くない箇所を推定するために，テキストの各文に出現する要素を行列で表現することによる，複数文からなる断片に対する局所的一貫性モデルを用いる研究を行った [4]．

これらの研究に対して本稿は機械学習を用いて順序推定を行うという違いがある．

岡崎らは，複数文書からの要約作成のために，複数の記事から抽出した文の順序を推定する研究を行った [5]．要約前の文章での文の順序も考慮して，複数の記事から抽出した文の順序を推定した．

Danushka らは複数文書からの要約作成のために文の順序推定の研究を行った [6]．文の順序推定には，時間的情報，内容の意味的近さ，要約前文章での文の順序な

どの情報を素性とした教師あり機械学習法を用いた．

これらの研究に対して本稿は要約前の文章の情報を利用しないという違いがある．要約前の文章の情報を利用せずに文章の順序を推定できれば，文章の順序が良くない文章の修正に役立つ．

# 3 問題設定と提案手法

## 3.1 問題設定

本研究での問題設定を以下に示す．記事のある箇所まで段落の順序が確定しており，それより後の箇所の段落の順序が不明であるとする．段落内の文は正しい順序であるとする．不明な箇所の先頭の 2 段落について，段落の順序を推定する．推定で用いることができる情報は，順序を推定する 2 段落とその 2 段落以前のその記事内の全ての段落とする．

## 3.2 提案手法

段落の順序を推定する 2 段落が順序付き (正順または逆順) で入力された場合，その順序が正解の順序と同じ順序か否かを教師あり機械学習で判定する．教師あり機械学習としては *SVM* を利用する．*SVM* には，*TinySVM*[7] を利用する．カーネル関数には 2 次の多項式カーネルを利用する．

学習データの作成方法は以下に示す．学習用の文章から連接する 2 段落を 1 組にして抜き出し，元の文章通りの順序 (正順) とその逆の順序 (逆順) の，2 つの問題を作成する．

順序の推定方法は以下に示す．順序を推定すべき段落対が順序付き (正順または逆順) で入力された場合，教師あり機械学習により，その順序が正しいかどうかを推定する．

## 3.3 提案手法で用いる素性

機械学習で用いられる識別用の情報のことを素性といい，機械学習は与えられたデータを用いて上手く識別できるような素性を学習する．本研究で用いる素性を表 1 に示す．素性は 2 段落のうちどちらに出現したかを区別して用いる．また 2 段落の両方の情報を用いる素性では，2 段落のうちの一方を 1 段落目にする場合，もう一方を 1 段落目にする場合の 2 種類とも素性として用い，この 2 種類は別の素性として区別して用いる．

単語や品詞の情報の取得には，形態素解析システムの *ChaSen*[8] を用いる．

### 3.3.1 a1:段落内に出現する品詞とその単語

素性 a1 で用いる品詞は，名詞，形容詞，形容動詞，動詞，副詞，連体詞，接続詞のみとする．また，単語を素性として用いる場合は a1 の品詞のみとする．

### 3.3.2 a2:段落内各文において，助詞「は」で文を括り，その前部・後部で出現する品詞とその単語

段落は複数の文から構成されるため，助詞「は」が多く出現する．これに対処するために段落を文ごとに区切る．各文に対して助詞「は」を含む場合，その助詞「は」を境にしてその文を前部・後部に分け，前部と後部についてそれぞれ異なる素性とする．

文に対して助詞「は」が 2 つ以上出現する場合は，初めに出現する「は」を境にして分ける．また，文中に 1 つも助詞「は」が出現しない場合は，全て後部と考えて素性とする．

表 1: 用いる素性

| 素性 | 説明 |
|---|---|
| a1 | 段落内に出現する品詞とその単語 |
| a2 | 段落内各文において，助詞「は」で文を区切り，その前部・後部で出現する品詞とその単語 |
| a3 | 段落内文頭に連体詞や接続詞が出現するか否か |
| a4 | 段落内に日付けが出現するか否か |
| a5 | 1 段落目と 2 段落目に出現する名詞が一致した数 |
| a6 | 1 段落目と 2 段落目に出現する名詞が一致した数を 2 段落目に出現する名詞の数で引いた数 |
| a7 | 素性 a6 の値と推定する段落を入れ替えた場合の a6 の値の二つの差 |
| a8 | 1 段落目に出現する名詞と 2 段落目の素性 a2 の前部に出現する名詞が一致した数 |
| a9 | 1 段落目に出現する名詞と 2 段落目の素性 a2 の前部に出現する名詞が一致した数を 2 段落目の a2 の前部に出現する名詞の数で引いた数 |
| a10 | 素性 a8 の値と推定する 2 段落を入れ替えた場合の a8 の値の二つの差 |
| a11 | 素性 a9 の値と推定する 2 段落を入れ替えた場合の a9 の値の二つの差 |
| a12 | 推定する 2 段落以前の段落と 1，2 段落目に出現する名詞が一致した数 |
| a13 | 推定する 2 段落以前の段落と 1，2 段落目に出現する名詞が一致した数を各段落に出現する名詞で引いた数 |
| a14 | 素性 a12 の値と推定する 2 段落を入れ替えた場合の a12 の値の二つの差 |
| a15 | 素性 a13 の値と推定する 2 段落を入れ替えた場合の a13 の値の二つの差 |
| a16 | 推定する 2 段落以前の段落に出現する名詞と 1，2 段落目の素性 a2 の前部に出現する名詞が一致した数 |
| a17 | 推定する 2 段落以前の段落に出現する名詞と 1，2 段落目の素性 a2 の前部に出現する名詞が一致した数を各段落目の a2 の前部に出現する名詞の数で引いた数 |
| a18 | 素性 a16 の値と推定する 2 段落を入れ替えた場合の a16 の値の二つの差 |
| a19 | 素性 a17 の値と推定する 2 段落を入れ替えた場合の a17 の値の二つの差 |
| a20 | 1 段落目と 2 段落目に出現する，推定する以前の段落に出現せず，かつ初めて出現する単語 (以下新規単語) の数の差 |
| a21 | 1 段落目に出現する新規単語とと 2 段落目に出現する新規単語の比率の差 |

### 3.3.3 a3:段落内文頭に連体詞や接続詞が出現するか否か

「この」や「それ」など連体詞が出現する場合は，以前に出現した単語を指し示している．また，「または」や「しかし」など接続詞が出現する場合は，文間の文脈の関係を示している．従って，連体詞や接続詞が文頭に出現する場合は以前に段落が存在すると考えられる．

### 3.3.4 a4:段落内に日付けが出現するか否か

注目される事柄が記載される場合は日付もその段落に書かれることが多く，注目される事柄は記事内の最初の方に書かれることが多いため，日付が出現する段落は前の方に記載される傾向がある．その傾向を使うためにこの情報を素性として用いる．

### 3.3.5 a5:1 段落目と 2 段落目に出現する名詞が一致した数

段落内には名詞が多く出現する．このことから名詞の一致に注目する素性を作成する．各段落に出現する名詞が一致した数を求め，その値が，0 以上，1 以上，2 以上を最大値 10 まで，0 以上 2 未満，2 以上 4 未満，を最大値 8 までの範囲で場合わけしたものを素性とする．また，素性 a8 は素性 a2 の前部に出現する名詞を用いることのみ a5 と異なるだけであるため，説明を省略する．

#### 3.3.6 a6:1 段落目と 2 段落目に出現する名詞が一致した数を 2 段落目に出現する名詞の数で引いた数

2 段落目に出現する名詞で，1 段落目の名詞と一致しなかった名詞の数である．素性 a9 も同様に前部に出現する名詞に限ることのみ a6 と異なるだけであるため説明を省略する．

#### 3.3.7 a7:素性 a6 の値と推定する段落を入れ替えた場合の a6 の値の二つの差

段落順を推定する 2 段落のうちの 1 段落目を L，2 段落目を R とする場合，「L → R」という順の 2 段落で名詞の一致数 A と逆の「R → L」という順の 2 段落で名詞の一致数 B を求め，A-B の値から，以下の条件に当てはまれば素性とする．値が 0 以上，0 未満かや -2 以上 0 未満，0 以上 2 未満，2 以上 4 未満，2 ずつの増減で最大値 8 で最小値-8 の範囲で場合わけしたものを素性とする．

#### 3.3.8 a12:推定する 2 段落以前の段落と推定する 2 段落の 1，2 段落目に出現する名詞が一致した数

連接する段落間の情報は似通っている方が文章の順序としては良い．このことから，推定する 2 段落が存在する記事内以前のすべての段落との名詞の一致数が多い段落が先に推定するように a12 の素性を設ける．素性 a12 から a19 までは前の素性を組み合わせることにより，情報を取得できるので，説明を省略する．

#### 3.3.9 a20:1 段落目と 2 段落目に出現する新規単語の数の差

まず，推定する 2 段落以前の文章に対して，1 段落目 (または 2 段落目) に出現する新規単語の数を A(または B) として求める．A-B を行い，その値が 0 未満か，0 より大か (超過) を示す素性を付与する．ここで用いる単語は a1 で用いる品詞のもののみとする．

#### 3.3.10 a21:1 段落目と 2 段落目に出現する新規単語の比率の差

推定する 2 段落以前の文章に対して，1 段落目 (または 2 段落目) に出現する新規単語の割合を A(または B) として求める．新規単語の割合は，新規単語の数を，その段落に出現する単語全ての数で割った値である．A-B を行い，新規単語の出現する比率が多い段落を後ろの方の順となるように，新規単語が少ない段落タグの素性を付与する．ここで用いる単語も a1 で用いる品詞のもののみとする．

## 4 比較手法

連接する 2 段落の情報は似通う．これにより，以下の手法を比較手法として用いる．推定する 2 段落以前の段落に出現する名詞と，推定する 2 段落それぞれに出現する名詞との一致数を求めて，一致数が大きい方の段落を前となる順序とする．この手法を比較手法 1 と呼ぶ．

また，新規単語は先頭段落を除き，後の段落の方につれて多く出現する場合があることから，以下の手法も比較手法として用いる．推定する 2 段落以前の段落に対する新規単語の比率を推定する 2 段落それぞれで求め，比率の小さい方の段落を前となる順序にする．この手法を比較手法 2 と呼ぶ．ここで用いる単語は a1 で用いる品詞のもののみとする．

本稿では，以上の比較手法の性能を提案手法の性能と比較する．

## 5 実験

### 5.1 実験条件

教師あり機械学習に用いる学習用文章には，毎日新聞 1992 年 7 月の 1 ヶ月分の記事を用いる．

実験で用いる 2 段落の組には，以下の 2 種類の場合を考慮して作成する．記事内の最初の 2 段落のみを用いて作成する場合 (Case1)，先頭段落も含む段落内全ての連接する 2 段落を用いて作成する場合 (Case2) とする．Case1 は先頭 2 段落のみを用いるため，比較手法の 4 節で挙げた比較手法は用いることができない．

Case1 は，先頭の段落対であり，推定する 2 段落以前の段落が存在しないので，推定する 2 段落以前の段落情報を用いる素性 (a12 から a21 まで) は用いない．Case2 は中間の段落も用いるので，接続詞や連体詞で順序を推定することが難しい．ゆえに，Case2 の場合は a3 を用いない．

Case1 に用いる学習データの 2 段落対の組数は 1,550 組，Case2 の組数は 29,434 組である．

### 5.2 提案手法と比較手法の比較実験

テストデータは毎日新聞記事 1992 年 8 月 1 日の 1 日分から作成する．Case1 での段落対の組数は 418 組であり，Case2 での段落対の組数は 3,146 組である．表 2 に提案手法と比較手法の正解率を示す．

表 2 の提案手法と比較手法を比較すると，Case1 での提案手法 (0.8517) は比較手法はないが約 8 割という高い正解率をであり，Case2 での提案手法 (0.5976) は比較手法 1(0.5277) や比較手法 2(0.5257) より正解率が高いことが分かる．

表 2: 提案手法と比較手法との正解率

| | 提案手法 | 比較手法 | |
|---|---|---|---|
| | | 比較手法 1 | 比較手法 2 |
| Case1 | 0.8517 | | |
| Case2 | 0.5976 | 0.5277 | 0.5257 |

### 5.3 人手との比較実験

提案手法と比較手法の性能を，人手による段落の順序推定の性能と比較する．人手による推定は被験者 2 名で別々に行う．

Case1 の場合は毎日新聞記事 1993 年 6 月の 1 ヶ月分の記事から，Case2 の場合は同年 7 月の 1 ヶ月分から，ランダムに 2 段落対 1 組を 50 組を抜き出し，これらをテストデータとして用いる．

表 3 に提案手法と比較手法と被験者の正解率を示す．平均は，被験者の正解率の平均を示す．

表 3 の提案手法と比較手法と被験者を比較すると，Case1 では提案手法 (0.88) が被験者の平均 (0.88) の性能と同等であることから，Case1 での提案手法は人間と同程度の性能を持つことが分かる．また，Case2 を見ると，提案手法 (0.60) は比較手法 1(0.56)，比較手法

2(0.54) より高いが，被験者の平均 (0.66) よりも低い．Case2 を全体的に見ると，提案手法だけでなく人手も六割と低い正解率であるため，Case2 のような連接している段落の順序推定は難しいと思われる．

表 3: 提案手法・比較手法・人手の順序推定の正解率

| | 提案手法 | 比較手法 | | 被験者 | | |
|---|---|---|---|---|---|---|
| | | 比較手法 1 | 比較手法 2 | A | B | 平均 |
| Case1 | 0.88 | | | 0.92 | 0.84 | 0.88 |
| Case2 | 0.60 | 0.56 | 0.54 | 0.68 | 0.64 | 0.66 |

### 5.4 *SVM* の分離平面からの距離を利用した素性分析

本稿で用いた素性のうち，段落の順序推定に有用な素性を確認するために，*SVM* の分離平面に基づく素性分析を行う．学習データで用いた個々の素性 1 個ずつを持つ事例を作成し，その事例を *SVM* で分類する．*SVM* で分類する際にその事例と分離平面の距離が算出される．距離の大きい素性が順序推定に重要な素性である．

上記の素性分析を行った結果，Case1 の場合，有用な素性の上位には「この」(この素性があればこの素性がある方の段落が後方となる学習をしていた) や素性 a3,a4 があった．最も有用な素性は素性 a2 の前部の場合の「日」(この素性があればこの素性がある方の段落が前方となる学習をしていた) となった．Case2 の場合，有用な素性の上位には素性 a11，a13，a21 があった．また，Case1 で有用な素性の上位であった a3 の元となる連体詞や接続詞は Case2 では上位になかったが，「日」を素性とするものが Case2 でも最も有用な素性となっていた．

### 5.5 文の順序推定と段落の順序推定の比較

本節では文の順序推定の研究を行った林ら [2] の結果と比較することで，文の順序推定と段落の順序推定の違いを考察する．本稿は林らの Case1，Case2 に相当する実験を行っている．林らの研究での Case1 は，段落内の最初の 2 文の順序を推定するものであり，推定に利用する情報はその 2 文のみである．林らの研究での Case2 は，段落内の連接する 2 文の順序を推定するものであり，推定に利用する情報はその 2 文の存在する段落内のその 2 文が出現するまでの文章である．おおよそ本稿の記事と段落が，林らの研究の段落と文に相当する．林らの結果を表 4 に示す．林らも本稿と同様に機械学習を利用している．ただし，用いる素性は本稿と異なる．林らも人手による順序推定も行っており，その結果も表に示している．

林らと本稿を比較すると，Case1 では本稿の提案手法の方が正解率が高い．Case1 は先頭における文/段落の順序推定であり，前方の情報を処理に用いない．その先頭の二つの文/段落のみで順序を推定する．この場合は，順序を推定する文/段落の箇所の情報のみで推定するために，文/段落に情報が多いほど推定しやすくなると思われ，これにより段落を扱う本稿の提案手法の方が正解率が高かったと思われる．

Case1 での人手の正解率を文と段落で比較すると段落の方が高い．この結果も上述の機械学習による順序推定と同様な傾向である．Case1 では段落の方が問題が簡単であることがわかる．

また，Case2 では林らの提案手法の方が正解率が高い．Case2 では文章の途中における文/段落の順序推定を行うため，前方の文章の情報を処理に用いる．Case2 では前方の文章との関係を利用する処理が重要となる．段落は文に比べて，段落内で話が完結してしまう可能性があるため，前方の文章との関係がそれほど順序推定のヒントにならないことが多い．このため，林らの提案手法の方が正解率が高かったものと思われる．

Case2 での人手の正解率を文と段落で比較すると文の方が高い．この結果も上述の機械学習による順序推定と同様な傾向である．Case2 では文の方が問題が簡単であることがわかる．

表 4: 林らの提案手法と人手の順序推定の正解率

| | 林らの提案手法 | 人手の順序推定 |
|---|---|---|
| Case1 | 0.79 | 0.82 |
| Case2 | 0.67 | 0.87 |

## 6 おわりに

本稿では段落の順序推定に教師あり機械学習を用いる手法を提案した．段落の順序を推定する実験において，記事先頭 2 段落の順序推定を行った場合提案手法は 0.85 という高い正解率を得た．人手による順序推定と同等の正解率であった．素性分析の結果，先頭 2 段落での順序推定には連体詞や接続詞，日付けの情報が有効だとわかった．また，連接した 2 段落での順序推定では，提案手法は約 6 割という正解率であった．前方の文章との名詞の一致数が大きい方を前方とするベースライン手法よりは高い正解率であった．素性分析の結果，連接した 2 段落での順序推定には新規単語の割合や日付けの情報が有効だとわかった．連接した 2 段落は各段落の情報がその段落内で完結している可能性があることから，段落の順序推定が他の場合より難しいことがわかった．文の順序推定と段落の順序推定の比較を行い，違いを明らかにした．

## 謝辞

本稿は科研費 (23500178) の助成を受けたものである．

## 参考文献

[1] 内元 清貴, 村田 真樹, 馬 青, 関根 聡, 井佐原 均: “コーパスからの語順の学習”, 情報処理学会研究報告, 自然言語処理研究会報告, 2000(11), pp. 55-62, 2000.
[2] 林 裕哉, 村田 真樹, 徳久 雅人: “教師あり機械学習を用いた文の順序推定”, 言語処理学会 第 18 回年次大会 発表論文集, pp. 239-242, 2012.
[3] Mirella Lapata: “Probablistic text structuring: Experiments with sentence ordering”, In Proceeding of the 41st Meeting of the Association of Computational Linguistics, pp. 545-552, 2003.
[4] 横野 光, 奥村 学: “テキストの断片に対する局所的一貫性モデル”, 情報処理学会研究報告, 自然言語処理研究会報告 2010, NL-199 17, pp. 191-194, 2010.
[5] 岡崎 直観, 石塚 満: “複数の新聞記事から抽出した文の並び順の検討”, 人工知能学会 第 18 回全国大会 発表論文集, pp. 191-194, 2004.
[6] Danushka Bollegala, Naoaki Okazaki, Mitsuru Ishizuka: “A bottom-up approach to sentence ordering for multi-document summarization”, Information Processing & Management, Vol. 46, No. 1, pp. 89-109, 2010.
[7] TinySvm : http://chasen.org/taku/software/TinySVM/
[8] ChaSen : http://chasen-leagacy.sourceforge.jp/