概要

本研究では文章を適切な順序に並べ替えるために，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0.88の正解率となり，人手による順序推定 (0.88) と同等の正解率であった．連接した2段落での順序推定を行った場合，提案手法 (0.60) は人手による順序推定 (0.66) より低いが比較手法 (0.56) より高い正解率であった．ここで用いた比較手法とは，前方の文章との名詞の一致数を順序推定を行う段落それぞれ求め，一致数がより大きい方の段落を前に推定するという手法である．

また，文の順序推定を扱った林らの研究結果に基づき，文と段落の順序推定結果を比較した．その結果，記事内先頭2（文・段落）対の順序を推定することについては，文より段落の方が推定しやすく，記事内連接2対の順序を推定することについては，段落より文の方が推定しづらいことがわかった．先頭2対の順序を推定する場合では以前の情報がないため，扱う情報が推定する2段落のみとなることから，推定情報が多い段落の方が推定がしやすくなると思われる．連接するあらゆる2対の順序を推定する場合では，段落は各段落内部で話題が完結し前方の文章との関係が小さいため，文より段落の方が推定しづらいと思われる．

目 次

表 目 次

図 目 次

第1章　はじめに

我々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説明なく用いることや，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，指示語を多く用いること，文章の順番が良くないことなど，様々な原因が存在する．これらの原因の処理を行うことにより，読みやすい文章となる．本論文では，上述の問題のうち，文章の順番を対処する．

文章の順番が良くないために読みづらい文章となっている場合は，文章を適切な順序に並べ替える必要がある．文章の順序推定に確率モデルを用いた手法 [1, 2] があるが，本論文では 2 値分類に秀でた分類器である “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．

教師あり機械学習を用いた文章の順序推定として，内元ら [3] や林ら [4] 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ている．ゆえに，本論文では段落の順序推定を行う．

本論文の特徴を以下に整理する．

- 段落の順序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られる可能性がある．
- 記事の最初の 2 段落における順序推定では，提案手法で (0.85) という高い正解率であった．この正解率は人手による順序推定の正解率と同等であった．
- 記事内の連接した 2 段落における順序推定では，提案手法で (0.60) という正解率であった．この正解率は人手による順序推定の正解率には劣るものの，“推定する 2 段落のうち前方の文章との名詞の一致した数が大きい段落を前方に推定する” 比較手法より高い正解率であった．
- 文と段落の順序推定の結果に対し比較を行い，段落の特徴を明らかにした．

第2章　関連研究

順序推定に用いる手法と順序推定の対象，用いる文章の３つの点に基づいて次節以降で関連研究を整理する．

2.1　順序推定に用いる手法が本研究と異なる関連研究

Lapata[1] 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推定する研究を行った．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 [2] 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．

これらの研究に対して，本論文では推定対象が文ではなく “段落” であり，また，確率手法ではなく “教師あり機械学習” を用いて順序推定を行うという違いがある．教師あり機械学習は 2 値分類に長けた分類器であり，また素性を簡単に，かつ大量に組み込むことができるため，有用な素性の発見が推定の性能向上となる．

2.2　順序推定の推定対象が本研究と異なる関連研究

内元ら [3] 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．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た．

林ら [4] 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．新聞記事から 2 文 1 組で抜き出し，その 2 文対から元の順の文（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推定を行うというものである．教師あり機械学習に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的に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，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 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，本論文は “段落” の順序推定を扱うという違いがある．

2.3　用いる文章が本研究と異なる関連研究

岡崎ら [5] 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．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 [6] は複数文書からの要約作成のために，文の順序を推定する研究を行った．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を用いた．

これらの研究に対して，本論文では推定対象が “段落” である点もあるが，ここでは “要約前の文章情報を用いない” という違いがある．要約前の文章の情報を用い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と思われる．

第3章　問題設定

本論文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，その後の箇所の段落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．推定で用いることのできる情報は，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順序が確定している全ての段落とする．

以上の問題設定の概念図を表3.1に示す．

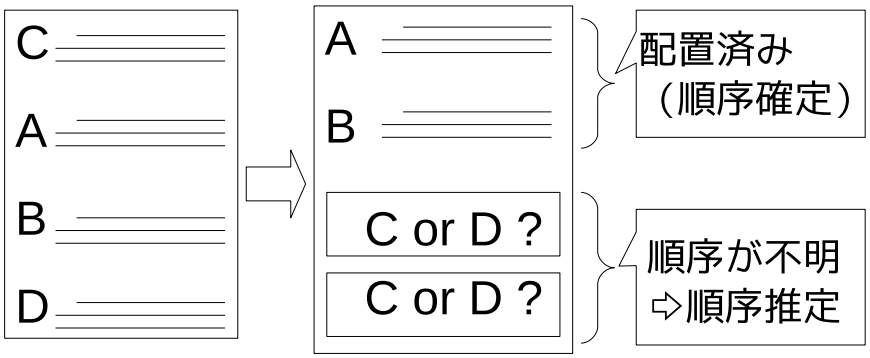


図 3.1: 問題設定の概念図

図3.1の場合，段落AとBは順序が確定しており，その後の段落C，Dの順序が不明である．この段落C，Dについて順序推定を行う．推定に用いられる情報は順序が確定している段落A，Bと推定箇所の段落C，Dである．

第4章　提案手法

4.1　2段落の順序推定方法

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．教師あり機械学習には SVM[1] を用いる．カーネル関数には2次の多公式をカーネルを利用する．

4.2　教師あり機械学習法

教師あり機械学習について説明する．図4.1のように，学習器に，扱う問題とその問題に対する出力といった入出力対の事例が複数与えられた学習データを学習させておき，問題のみのテストデータが新たに入力されたとき，学習データを基にテストデータに対する正しい出力を行うことが目的である．

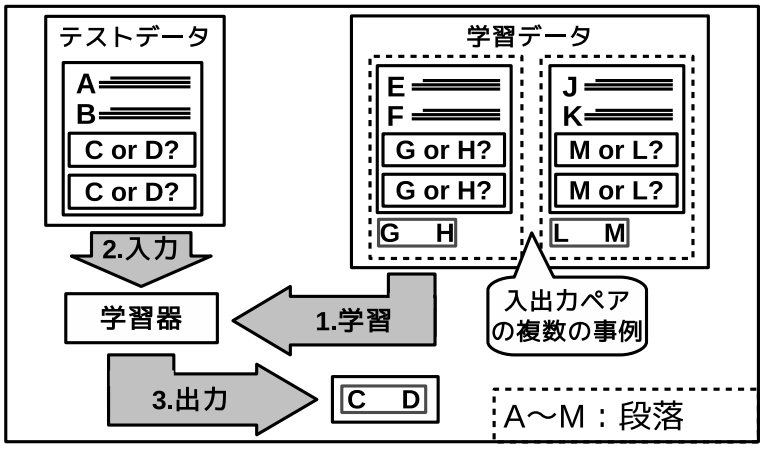


図 4.1: 教師あり機械学習の概念図

[1]具体的に SVM には $TinySVM$[7] を用いる．

4.3 SVM法

SVMは，空間を超平面で分割することにより2つの分類からなるデータを分類する手法である [8]．このとき，2つの分類が正例と負例からなるものとすると，学習データにおける正例と負例の間隔 (マージン) が大きいもの (図 4.2 参照[2]) ほどオープンデータで誤った分類をする可能性が低いと考えられ，このマージンを最大にする超平面を求めそれを用いて分類を行う．基本的には上記のとおりであるが，通常，学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や，超平面の線形の部分を非線型にする拡張 (カーネル関数の導入) がなされたものが用いられる．この拡張された方法は，以下の識別関数を用いて分類することと等価であり，その識別関数の出力値が正か負かによって二つの分類を判別することができる．

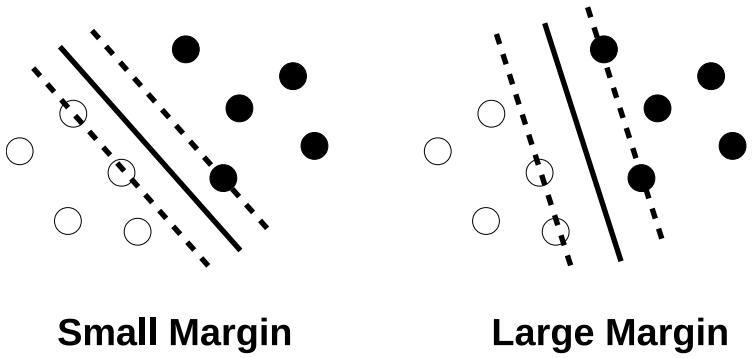


図 4.2: マージン最大化

$$
\begin{aligned}
f(\mathbf{x}) &= sgn\left(\sum_{i=1}^{l}\alpha_i y_i K(\mathbf{x}_i,\mathbf{x})+b\right) \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{(4.1)} \\
b &= -\frac{max_{i,y_i=-1}b_i + min_{i,y_i=1}b_i}{2} \\
b_i &= \sum_{j=1}^{l}\alpha_j y_j K(\mathbf{x}_j,\mathbf{x}_i)
\end{aligned}
$$

ただし，$\mathbf{x}$ は識別したい事例の文脈 (素性の集合) を，$\mathbf{x}_i$ と $y_i(i=1,...,l,y_i\in\{1,-1\})$

[2]図の白丸，黒丸は，正例，負例を意味し，実線は空間を分割する超平面を意味し，破線はマージン領域の境界を表す面を意味する．

は学習データの文脈と分類先を意味し，関数 sgn は，

$$sgn(x) = \begin{array}{ll} 1 & (x \geq 0) \\ -1 & (otherwise) \end{array} \quad \cdots\cdots\cdots\cdots\cdots\cdots (4.2)$$

であり，また，各 α_i は式 (4.4) と式 (4.5) の制約のもと式 (4.3) の $L(\alpha)$ を最大にする場合のものである．

$$L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \cdots\cdots\cdots\cdots\cdots\cdots (4.3)$$

$$0 \leq \alpha_i \leq C \ (i = 1, ..., l) \quad \cdots\cdots\cdots\cdots\cdots\cdots (4.4)$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \quad \cdots\cdots\cdots\cdots\cdots\cdots (4.5)$$

また，関数 K はカーネル関数と呼ばれ，様々なものが用いられるが本論文では以下の多項式のものを用いる．

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^d \quad \cdots\cdots\cdots\cdots\cdots\cdots (4.6)$$

C, d は実験的に設定される定数である．本論文ではすべての実験を通して C,d はそれぞれ 1 と 2 に固定した．ここで，$\alpha_i > 0$ となる $\mathbf{x}_i$ は，サポートベクトルと呼ばれ，通常，式 (4.1) の和をとっている部分はこの事例のみを用いて計算される．つまり，実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない．

4.4 データ作成

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 段落対を 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．その後，段落内の情報から各素性を求め (素性については 5 章参照)，学習データを作成する．

テストデータも同様に，テスト用の文章から作成する．学習データ同様の処理を施し，テストデータが作成される．テストデータにも順序が付与されるが，SVM の出力が正しいかを比較するために用いる．

第5章　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．本論文で用いる素性を表 5.1 に示す．素性は推定する 2 段落のうちどちらに出現したかを区別して用いる．品詞や単語の情報の取得には形態素解析システムの *ChaSen*[9] を用いる．

各素性の詳細な説明を表 5.1 に示す．

表 5.1: 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 2 段落を入れ替えた場合の a6 の 2 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 2 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 2 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の数で引いた数
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 2 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 2 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 2 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 2 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 2 段落目に出現する新規単語の比率の差

a1：段落内に出現する品詞とその単語

素性 a1 は段落内に出現する品詞とその単語を素性としている．ここで用いられる品詞は，名詞，形容詞，形容動詞，動詞，副詞，連体詞，接続詞とする．

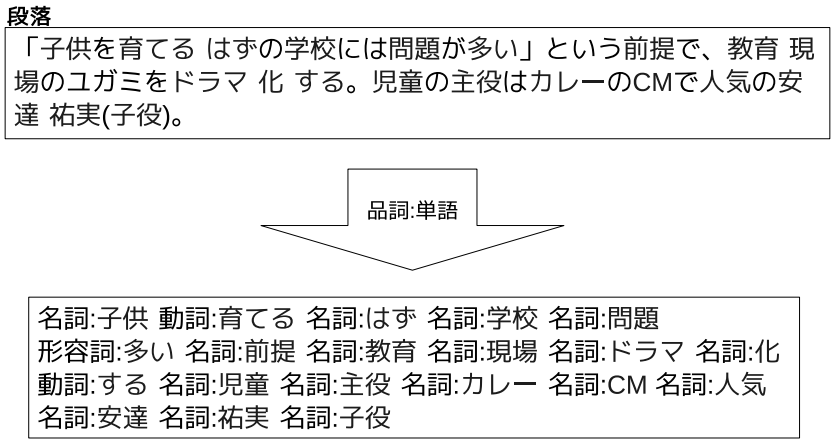


図 5.1: a1 の説明図

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

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・後部の 2 つに分け，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 2 つに分ける．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．

「には」や「では」のような「助詞＋ “は”」の場合の「は」は，助詞「は」として考慮しない（区切らない）．

以下の例では，区切るまでを示す.（出現する品詞とその単語に整理するのは素性 a1 の例参照）

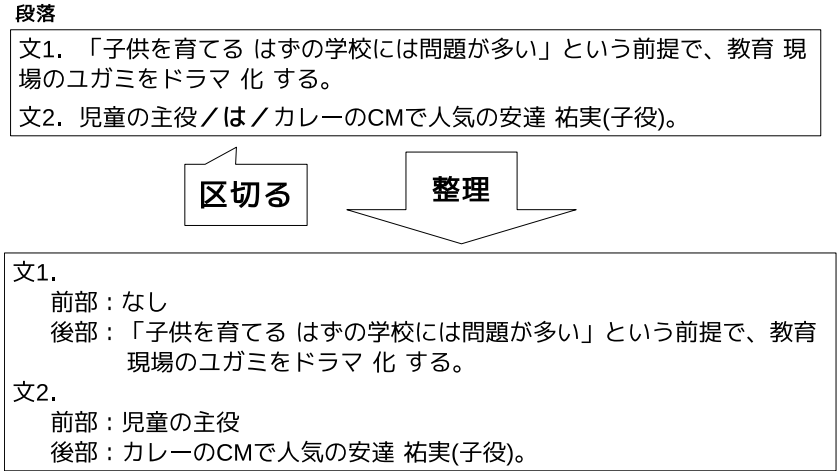


図 5.2: a2 の説明図

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「その」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，文章間における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．

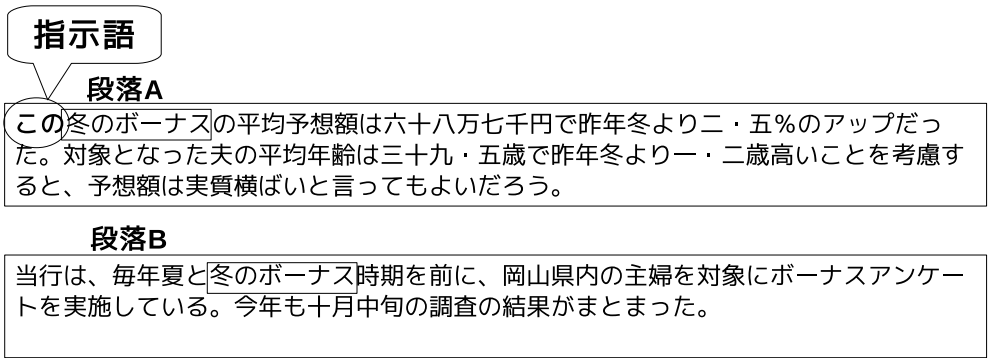


図 5.3: a3 の説明図

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け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，日付けが出現する段落は前の方に記載される傾向がある．この素性はその傾向のうち,「日」に注目している．

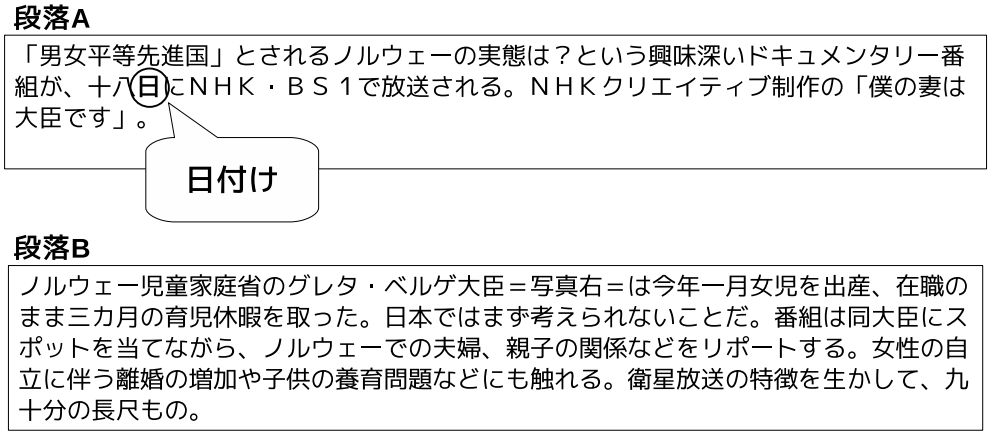


図 5.4: a4 の説明図

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数に着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，3以上を最大値10まで，0以上2未満，2以上4未満，4以上6未満を2ずつ増加し最大値8までの範囲で場合分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，素性a5と異なるだけであるため説明を省略する．

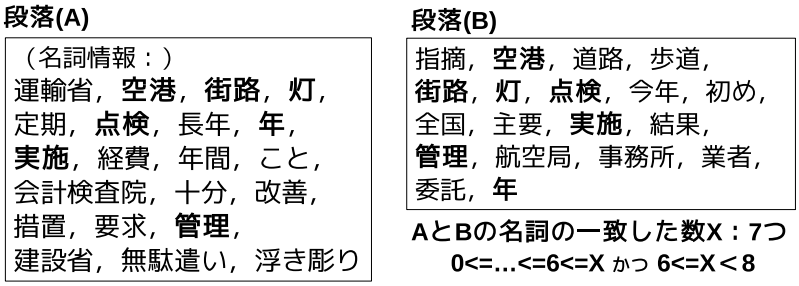


図 5.5: a5の説明図

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

素性a6は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のことである．その値を求め，素性a5同様の場合分けしたものを素性とする．素性a9も同様に前部に出現する名詞に限ることのみ，素性a6と異なるだけであるため説明を省略する．

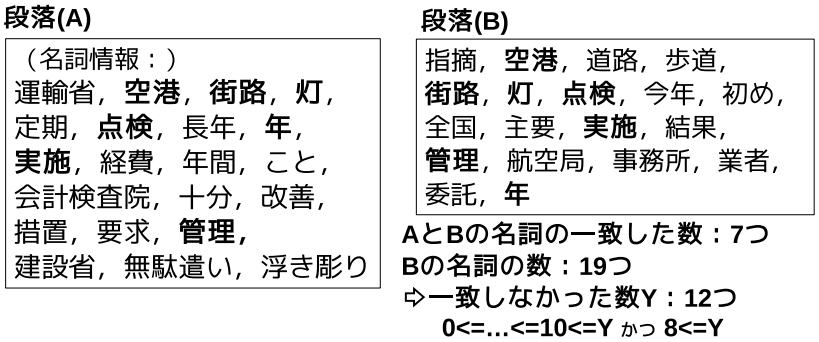


図 5.6: a6の説明図

a7：素性 a6 の値と推定する 2 段落を入れ替えた場合の a6 の値の 2 つの差

段落の順序を推定する 2 段落のうちの 1 段落目を A，2 段落目を B とする場合,「A → B」という順序の 2 段落で B の全名詞の内の一致しなかった X と，逆の「B → A」という順序の 2 段落で A の全名詞の内の一致しなかった数 Y を求める．X-Y の値から，値が 0 以上，0 未満かや-4 以上-2 未満，-2 以上 0 未満，0 以上 2 未満を 2 ずつ増減し最大値 8 最小値-8 の範囲で場合分けしたものを素性とする．素性 10，11 も同様であるため，省略する．

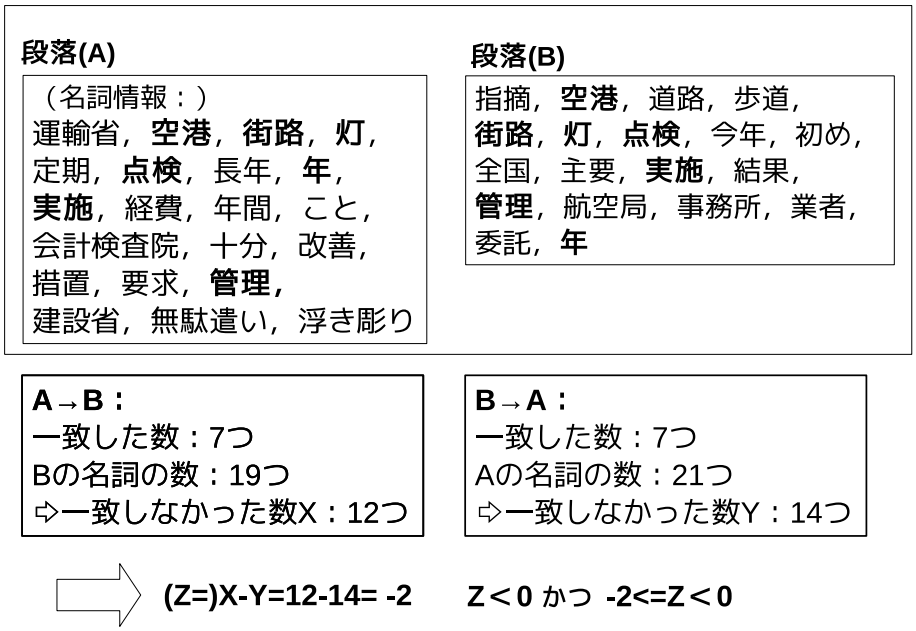


図 5.7: a7 の説明図

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

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全ての段落との名詞の一致した数が多い段落が前方に推定するように a12 の素性を設ける．素性 12 から素性 a19 までは前の素性を組み合わせることにより，素性を取得できるため，説明を省略する．

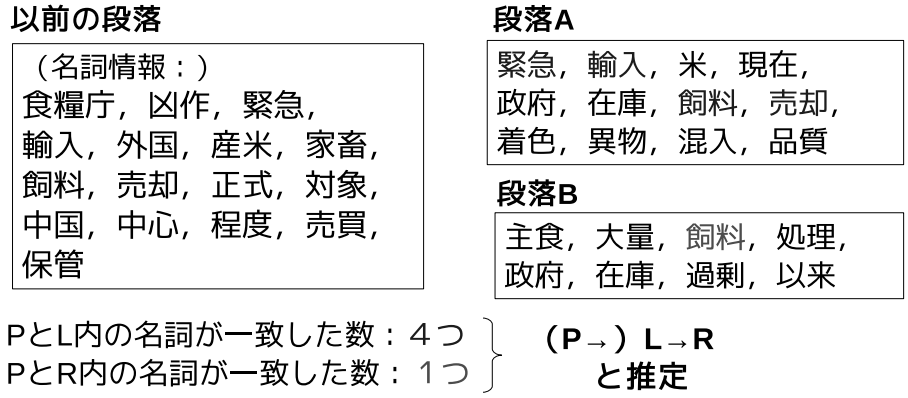


図 5.8: a12 の説明図

a20：1段落目と2段落目に出現する新規単語の数の差

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数X（またはY）を求める．X-Yを行い，その値が0未満か，0より大（超過）かを示す素性を付与する．ここで用いる品詞は素性a1同様，名詞，形容詞，形容動詞，動詞，副詞，連体詞，接続詞とする．

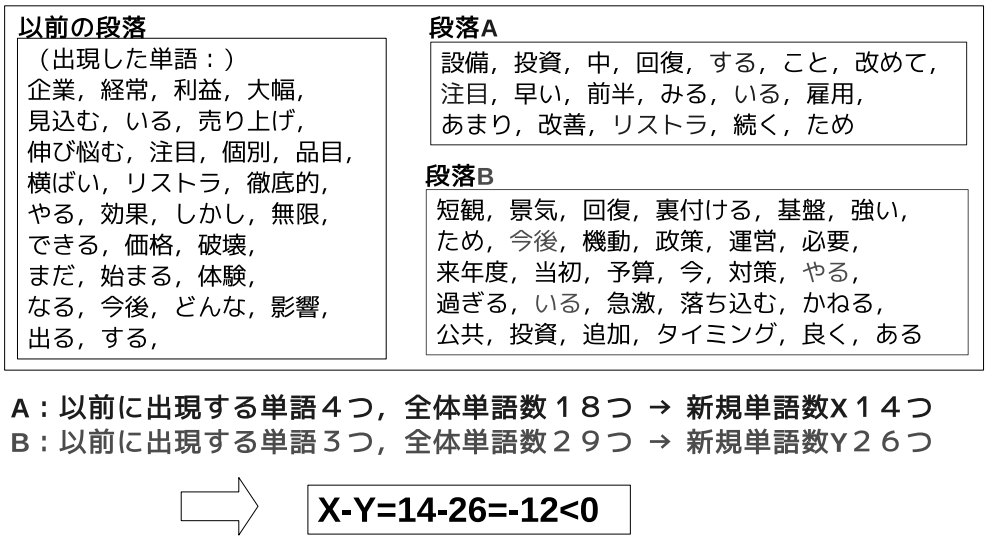


図 5.9: a20の説明図

a21：1段落目と2段落目に出現する新規単語の比率の差

新規単語の比率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．素性 a20 同様に X-Y を求め，新規単語の出現する比率が多い段落を後方に推定するように，新規単語が少ない段落タグの素性を付与する．ここで用いる品詞も素性 a20 同様のものとする．

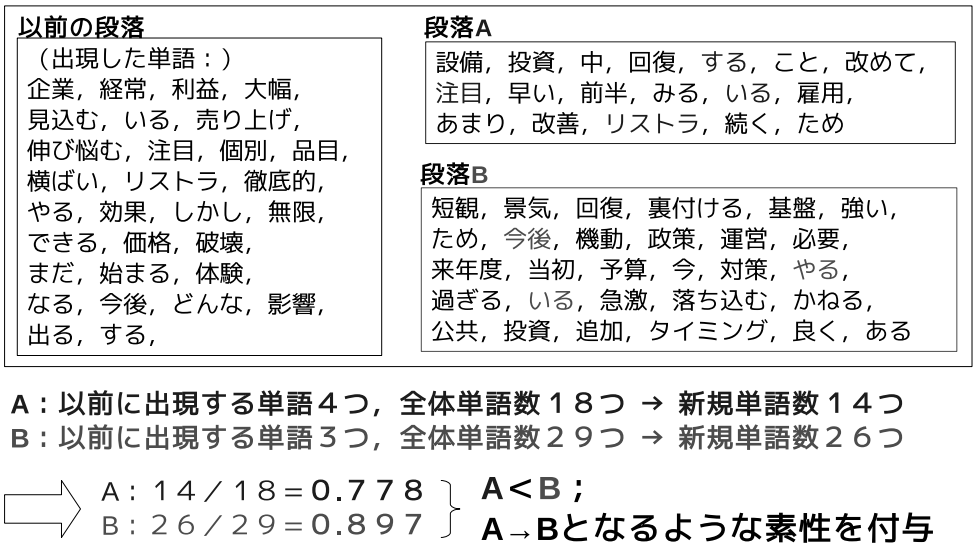


図 5.10: a21 の説明図

第6章　比較手法

提案手法の有効性確認のために以下の手法を用いて比較実験を行う．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．

比較手法1

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した数を求めて，一致した数が大きい方の段落を前方となる順序とする．

また，新規単語は先頭段落を除き，後方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．

比較手法2

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方となる順序とする．

本論文では，以上の比較手法の性能を提案手法の性能と比較する．（各比較手法の構想図は以下の素性の説明図参照；比較手法1：図5.8，比較手法2：図5.10）

第7章　実験

7.1　実験条件

実験で順序推定を行う 2 段落対の組には 2 種類の場合を考慮して作成する.

1．記事内の最初の 2 段落のみの対の順序を推定する場合（以下先頭 2 段落対，Case1 とも表記）

2．記事内のあらゆる連接する 2 段落の対の順序を推定する場合（以下連接 2 段落対，Case2 とも表記）

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ヶ月分を用いる．各場合での学習データの 2 段落対の組数を表 7.1 に示す.

表 7.1: 先頭 2 段落対と連接 2 段落対に用いる学習データの 2 段落対の組数（組）

	先頭 2 段落対	連接 2 段落対
学習データ	1,550	29,434

また，2 種類の比較実験 (実験 1，実験 2) において先頭 2 段落対，連接 2 段落対での順序推定を各々行う.

実験 1． 提案手法と比較手法による順序推定の比較

実験 2． 提案手法と比較手法と人手による順序推定の比較

テストデータには，実験 1 の場合毎日新聞 1992 年 8 月 1 日の 1 日分を用いる．実験 2 の場合での，先頭 2 段落対の順序推定を行う場合毎日新聞 1993 年 6 月の 1ヶ月分，連接 2 段落対の順序推定を行う場合毎日新聞 1993 年 7 月の 1ヶ月分を用いる[1]．各実験に用いたテストデータの組数を表 7.2，表 7.3 に示す．

表 7.2: 実験 1 でのテストデータの 2 段落対の組数（組）

実験 1	先頭 2 段落対	連接 2 段落対
テストデータ	428	3,146

表 7.3: 実験 2 でのテストデータの 2 段落対の組数（組）

実験 2	先頭 2 段落対	連接 2 段落対
テストデータ	50	50

[1]テストデータ:ランダムに 2 段落 1 組を 50 組抽出

1．記事内の最初の 2 段落のみを用いる場合

先頭 2 段落対の場合では記事内の最初の 2 段落のみの対を用いて，2 段落 1 組を作成する．作成した組において，作成に用いた記事での元の順序（正順）とその逆順を学習データ，テストデータそれぞれ作成する．学習データには段落対に用いた順序をそれぞれの段落対に付与し，テストデータもまた学習データ同様各段落対に順序をそれぞれの段落対に付与するが，テストデータの場合は教師あり機械学習からの出力による推定結果との正誤の際に用いる．

先頭 2 段落対のみを用いるため，6 章で挙げた比較手法は用いることができない．また，先頭 2 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 段落以前の情報を用いる素性 (素性 a12 から a21 まで）も用いることができない．

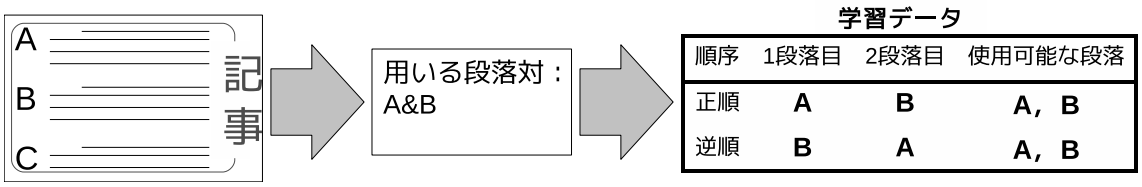


図 7.1: 先頭 2 段落対

図 7.1 の場合を例に挙げると，変数 A から C は段落であり，(A，B，C) で 1 記事とし，順序は上から順になるとする時，抽出される先頭 2 段落対は記事内の初めの 2 段落の対であるため，(A&B) になる．抽出された 2 段落対 (:A&B) から正順 (:A → B) と逆順 (:B → A) とそれぞれ入出力データとして学習データを作成する．また先頭 2 段落対を用いる場合なので，使用可能な段落は (A，B) となる．

2．記事内のあらゆる連接する 2 段落を用いる場合

連接 2 段落対の場合は記事内のあらゆる連接する 2 段落の対を用いて，2 段落 1 組を作成する．作成した組において，抽出した時の元の順序（正順）とその逆順を学習データ，テストデータそれぞれ作成する．学習データには段落対に用いた順序をそれぞれの段落対に付与し，テストデータもまた学習データ同様各段落対に順序をそれぞれの段落対に付与するが，テストデータの場合は教師あり機械学習からの出力による推定結果との正誤の際に用いる．

また，連接 2 段落対の場合は記事中の中間の段落も用いる．片方の段落に接続詞や連体詞が出現したとしてもこれらの品詞が指し示す順序を推定することは困難である．ゆえに，連接 2 段落対の場合は素性 a3 を用いない．

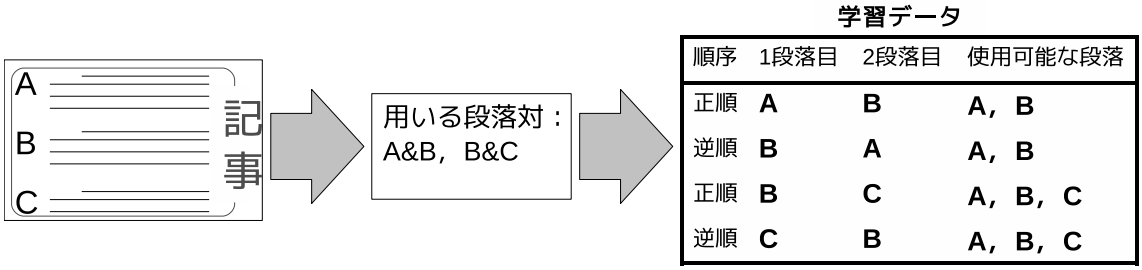


図 7.2: 連接 2 段落対

図 7.2 の場合を例に挙げると，図 7.1 の場合同様，変数 A から C は段落であり，(A，B，C) で 1 記事とし，順序は上から順になるとする時，抽出される連接 2 段落対は，(A&B) と (B&C) になる．抽出された 2 段落対 (:A&B) から正順 (:A → B) と逆順 (:B → A) とそれぞれ入出力データとして学習データを作成する．また連接 2 段落対を用いる場合なので，以前の段落も用いることができる．そのため，推定段落が (B，C) の時に使用可能な段落は (A，B，C) となる．

7.2 実験結果

7.2.1 実験 1．提案手法と比較手法 1 と比較手法 2 の順序推定の比較

提案手法と比較手法 1 と比較手法 2 の順序推定の正解率を表 7.4 に示す.

表 7.4: 提案手法と比較手法の順序推定の正解率

	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

表 7.4 から，Case1 では提案手法は約 8 割という高い正解率であり，Case2 では提案手法は比較手法 1，2 より高い正解率ことがわかる.

7.2.2 実験 2．提案手法と比較手法 1 と比較手法 2 と人手による順序推定の比較

さらに人による順序推定を加え比較を行う．提案手法と比較手法 1 と比較手法 2 と人手による推定の正解率を表 7.5 に示す．人手による推定では被験者 2 名 (A，B) により人手で順序を推定する．その 2 名の正解率の平均も表に示している.

表 7.5: 提案手法と比較手法 1 と比較手法 2 と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**

表 7.5 から，Case1 では提案手法が被験者の平均と同等であり人間と同等の性能であることがわかった．Case2 では提案手法，比較手法，被験者の平均ともに低い正解率であることがわかった.

7.2.3 推定結果の例

実際の推定結果の例を図 7.3 から図 7.8 に示す．図 7.3 は Case1 での提案手法×，人手推定○の考察例であり，図 7.4 は Case1 での提案手法×，人手推定×の考察例であ

り，図 7.5，図 7.6 は Case2 での提案手法○，比較手法×の考察例であり，図 7.7，図 7.8 は Case2 での提案手法×，比較手法○の考察例である．

Case1:提案手法×, 人手推定○ の実例 (正解順：C → D)

段落C

鶏卵と給食食材の卸業をしていたが、鶏卵業がスーパー業界に押され、一九八一年に新橋に居酒屋を持った。しかし、給食食材卸業との掛け持ちで過労のためダウンして入隠退院後の回復は早かったが、実は、**妻の好江さん**（５３）がみそ汁に客の食品会社の人が送ってきたスッポンのスープを入れていたことを後で知り「スッポンのおかげ」と感激。新橋と間もなく開店した日本橋の店でスッポン料理を始めた。

段落D

ＪＲ八王子駅の北西五百メートル足らずのところに、スッポンと地酒の店「鬼無里村（きなさむら）」がある。どこかで聞いたことがある名前だと思う人がいるにちがいない。そう、長野県北部に同名の村があり、七年前まで東京・新橋烏森口と日本橋の三越近くに同名の店があった。「村長」を名乗るオーナー、後藤昇さん（５９）はいう。

提案手法：D→C
（素性「段落内各文の助詞「は」の前部の単語」による影響？
「退院」，「後」等【断定不可】）
人手推定：C→D
（「妻の」＋「名前」：存在，かつ文頭から「妻」までに妻の「苗字」：出現否
→後方と推定）

図 7.3: 提案手法×, 人手推定○の考察例

Case1:提案手法×, 人手推定× の実例 (正解順：C → D)

段落C

自動改札機の普及で、定期券の受取口で取り間違えのトラブルが増えている。一人が取り間違えたことで後に続く人が執に間違え、一度に十人近くが"団体被害"を受けるケースも出ているという。鉄道総合技術研究所はセンサーで感知する機械を開発中だが、現在の機械では被害を完全に防止するのは無理、という。

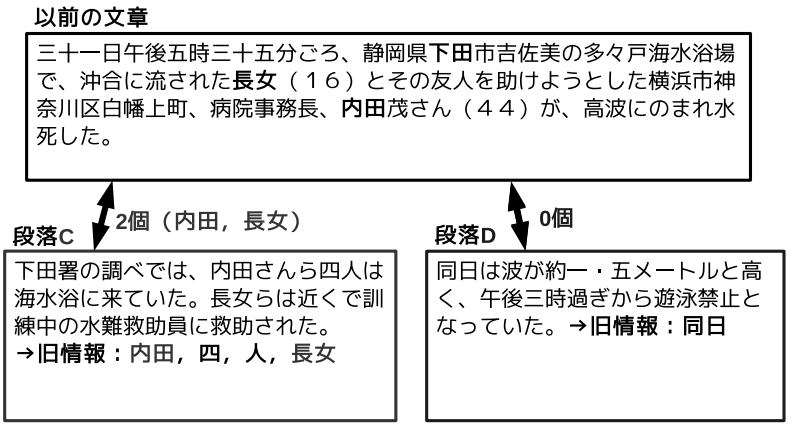
段落D

東京都千代田区の営団地下鉄竹橋駅で今月十四日夜、神奈川県内の会社員（５０）が自動改札機に入れた定期券を取り、しばらくして気づくと、他人のものだった。あわてて改札口に戻ったがすでに三、四人が改札口を通過しており、取り戻せなかった。

提案手法：D→C，人手推定：D→C
（先頭2段落対での順序推定：段落内に日付け→前方になる可能性：大
素性「段落内に日付けが含まれるか否か」）

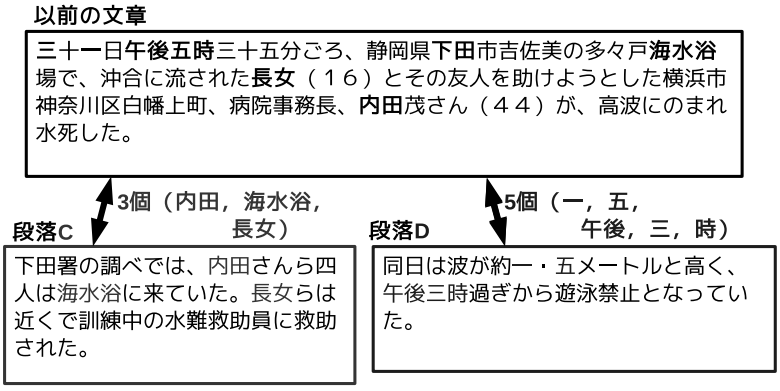
図 7.4: 提案手法×, 人手推定×の考察例

Case2:提案手法○, 比較手法× の実例 (正解順：C → D)



提案手法：C→D
（「段落内各文内の助詞「は」の前後部の単語(:a)」と「以前の文章内の名詞と推定する段落内の(a)の前部の名詞との一致数」の素性を使用
「は」の前部：既知【旧】情報，後部：未知【新】情報）

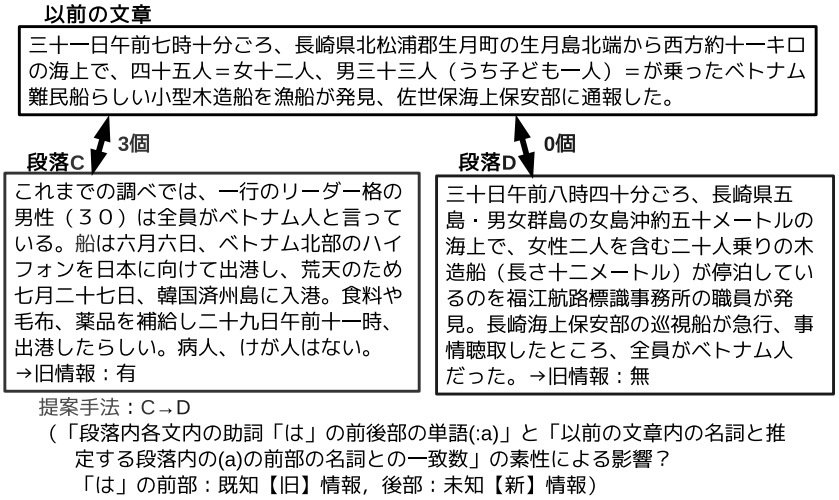
図 7.5: 提案手法○の考察例



比較手法：D→C
（以前の文章との単語の一致数→関係性のない単語でも一致）

図 7.6: 比較手法×の考察例

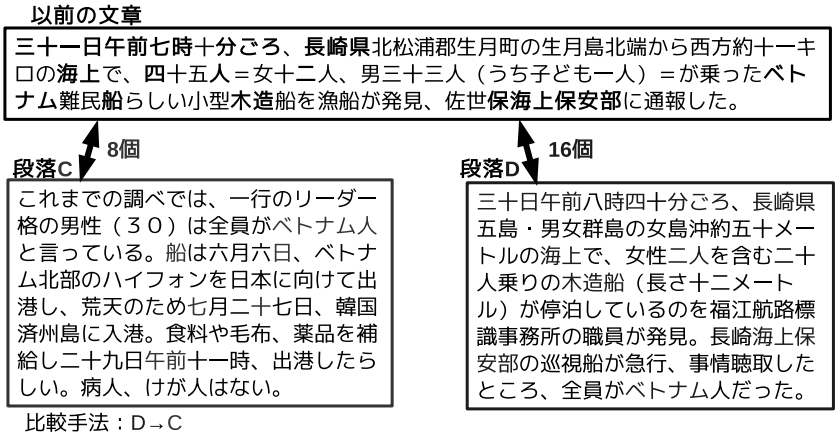
Case2:提案手法×, 比較手法○ の実例 (正解順：D → C)



提案手法：C→D

（「段落内各文内の助詞「は」の前後部の単語(:a)」と「以前の文章内の名詞と推定する段落内の(a)の前部の名詞との一致数」の素性による影響？
「は」の前部：既知【旧】情報，後部：未知【新】情報）

図 7.7: 提案手法×の考察例



比較手法：D→C

（以前の文章とC，Dそれぞれの名詞の一致数：CとDの一致数に大きな差）

図 7.8: 比較手法○の考察例

第8章　文の順序推定と段落の順序推定の比較

本章では文の順序推定の研究を行った林ら [4] 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．

本論文は林らの記事先頭 2 段落対の順序を推定する場合 (Case1) と記事あらゆる連接 2 段落対の順序を推定する場合 (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．おおよそ本論文の記事と段落が，林らの研究の段落と文に相当する．Case1 での各順序推定の正解率を表 8.1 に示し，Case2 での各順序推定の正解率を表 8.2 に示す．林らも本論文と同様に機械学習を利用している．ただし，用いる素性は本論文と異なる．林らも人手による順序推定も行っており，その結果も表 8.1，表 8.2 に示している．

表 8.1: Case1 での各順序推定の正解率

	提案手法	人手推定
文の順序推定	0.79	0.82
段落の順序推定	**0.88**	**0.88**

表 8.2: Case2 での各順序推定の正解率

	提案手法	人手推定
文の順序推定	**0.67**	**0.87**
段落の順序推定	0.60	0.66

林らと本論文を比較すると，Case1 では本論文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論文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．

段落の順序推定の場合

Case2 について，図を用いて説明する．段落の順序推定の場合を考える．図 8.1 の場合，段落 CD 間の名詞の一致数を求めるが，0 個なためこの段落間に関係性はない．仮に段落 C が以前の段落となり，段落 D と段落 D の次の段落との順序推定を行うならば，順序推定が困難であるだろう．

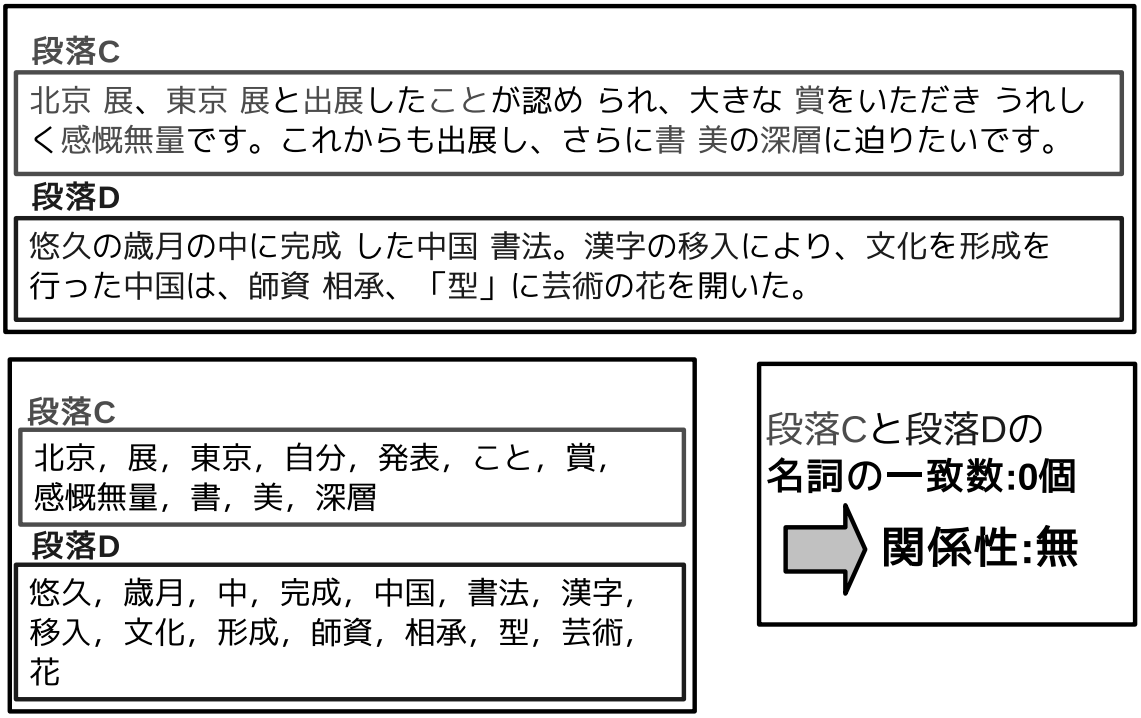


図 8.1: Case2 での段落の順序推定の考察例

文の順序推定の場合

文の順序推定の場合を考える．図 8.2 の場合，文 C_1 と文 D_1 間の名詞の一致数は 0 個となり，この文間の関係性は無い．しかし，(C_1 と C_2) や (D_1 や D_2) といったそれ以外の文間の名詞の一致数は 1 個あり，これらの文間には関係性が有ると言える．

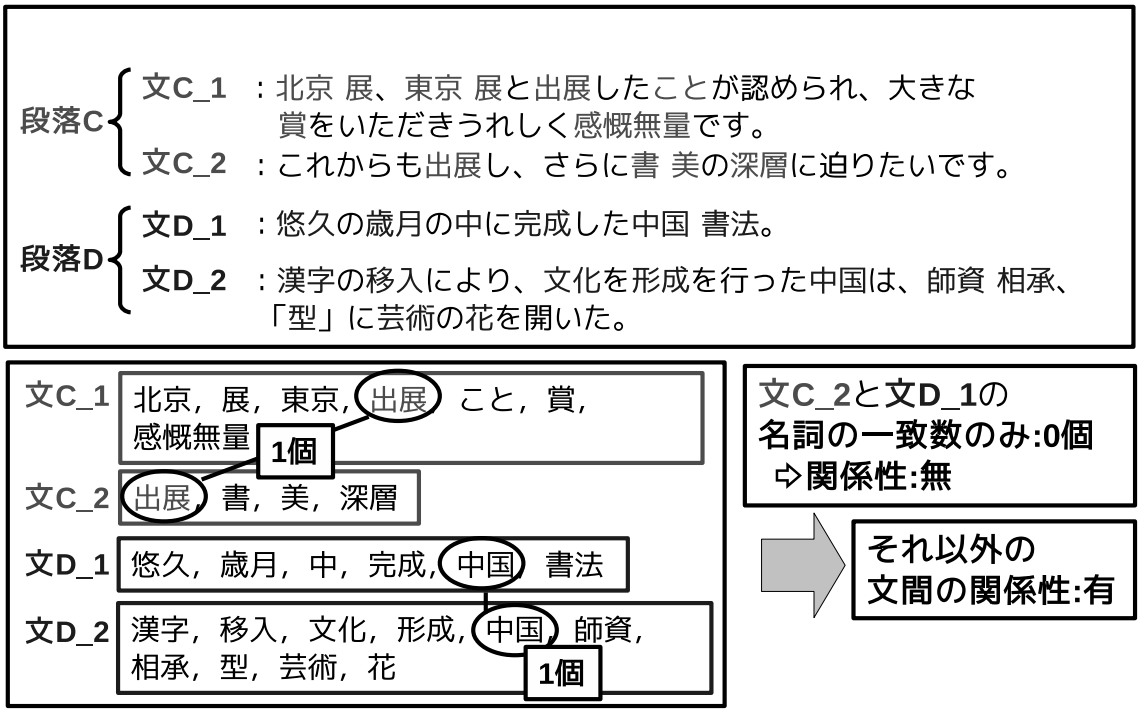


図 8.2: Case2 での文の順序推定の考察例

第9章　追加実験

本研究では以下の 4 点について追加実験を行った.

1． Case2 での素性 a3 を用いるか否かの順序推定の比較

2． 各 Case でのカーネル関数内の次数の増減による正解率の変化の調査

3． 素性の除去による素性分析

4． *SVM* の分離平面からの距離を利用した素性分析

9.1　Case2 での素性 a3 を用いるか否かの順序推定の比較

素性 a3 とは “段落内文頭に連体詞や接続詞が出現するか否か” というものである．本論文での Case2 の（推定対象にあらゆる連接する 2 段落対を用いる）場合に「新聞記事の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は困難である」という考察の上この素性 a3 を取り除いた．9.1 節では a3 を用いた場合と a3 を用いない場合を比較し，実際に a3 を用いないことが良かったのかを検証する.

9.1.1　実験条件

実験では，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 8 月 7 日の 1 日分を，テスト用文章には同新聞 1994 年 12 月 16 日の 1 日分の記事を用いる．実験に用いる学習データ，テストデータの組数を表 9.1 に示す.

表 9.1: 実験に用いる学習／テストデータの 2 段落対の組数（組）

学習データ	テストデータ
1,830	1,904

9.1.2 実験結果

実験結果は表 9.2 に示す．表 9.2 により，a3 を用いない場合の方が用いる場合より高い正解率となったため，Case2 では素性 a3 を用いないことが正しいとわかった．この節以降の追加実験には素性 a3 は用いないこととする．

表 9.2: 素性 a3 を用いる場合と用いない場合の Case2 での順序推定の正解率

a3 を	正解率
用いる場合	0.5783
用いない場合	0.5777

9.2 各 Case でのカーネル関数内の次数の増減による正解率の変化の調査

本研究では教師あり機械学習に *SVM*(4.3 節参照) を用いた．4.3 節でも述べた通り，*SVM* にはカーネル関数による拡張がなされている．本節では多項式のカーネル関数内の次数 d の増減による正解率の変化を調査することを目的とする．

9.2.1 実験条件

Case1（推定対象に先頭の 2 段落のみの対を用いる）の場合に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 8 月の 1ヶ月分を，テスト用文章には同新聞 1994 年 12 月 16 日の 1 日分を用いる．また，Case2 の場合での学習用文章には毎日新聞 1992 年 8 月 7 日の 1 日分を，テスト用文章には同新聞 1994 年 12 月 16 日の 1 日分の記事を用いる．各 Case の場合に用いる学習データ，テストデータの組数を表 9.3 に示す．

表 9.3: 各 Case に用いる学習／テストデータの 2 段落対の組数（組）

Case	学習データ	テストデータ
Case1	6,338	1,032
Case2	1,830	1,904

9.2.2 実験結果

実験に用いたカーネル関数の次数 d の値と各 Case それぞれの正解率を表 9.4 に示す．用いた SVM のツールでは，次数 $d = 4$ 以降の順序推定の判定が出力されなかった．

表 9.4: カーネル関数の次数 d の値と各 Case の正解率

Case	次数 d			
	1	2	3	4
Case1	0.8227	**0.8314**	0.8256	
Case2	0.5614	0.5783	**0.5914**	

表 9.4 より，Case1 では次数 $d = 2$ が (0.8314) と最も高い正解率 (2 番目との差:+0.0058) であり，Case2 では $d = 3$ が (0.5914) と最も高い正解率 (2 番目との差:+0.0131) となった．

結果のみを考慮すれば Case2 は $d = 3$ とするのが良いと思われるが，d が大きくなるにつれて，算出される時間（コスト）が掛かっている．$d = 1, 2$ に掛かるコストはそれほどでもないが，$d = 3$ の場合以降は膨大である．以上により，Case1，2 ともにカーネル関数の次数は $d = 2$ とした．

9.3 素性の除去による素性分析

素性の除去により段落の順序推定に役立つ有用な素性を調査する．

9.3.1 実験条件

素性の除去には，似たような素性をグループとしたものを用いる．除去する素性のグループを表 9.5 に示す（素性の詳細は 5 章参照）．実験に用いた学習用文章とテスト用文章は 9.2 節と同じものを用いる．

表 9.5: 除去する素性のグループ

分類	除去するグループ	Case1	Case2
1	a1,a2	○	○
2	a3,a4	○	○
3	a5,a6,a7	○	○
4	a8,a9,a10,a11	○	○
5	a12,a13,a14,a15		○
6	a16,a17,a18,a19		○
7	a20,a21		○
8	無	○	○

9.3.2 実験結果

除去した素性グループごとに各 Case の正解率を表 9.6 に示す．

表 9.6: 素性の除去による素性分析

分類	Case1	Case2
1	0.7393	0.5473
2	0.8256	0.5756
3	0.8304	0.5882
4	0.8217	0.5756
5		0.5788
6		0.5867
7		0.5767
8	0.8314	0.5783

表 9.6 より，分類 1 以外の増減は Case1，2 ともに他の分類に比べ増減が大きい．このため，有用な素性は分類 1 の a1，a2，すなわち段落内に出現する品詞とその単語を素性と表現する場合が有用であるといえる．

9.4 *SVM*の分離平面からの距離を利用した素性分析

*SVM*の分離平面に基づく素性分析により段落の順序推定に役立つ有用な素性を調査する．

9.4.1 実験条件

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とする．この節に用いる学習データは 7.2.1 節に用いたデータを使用する．

9.4.2 実験結果

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．

第10章　おわりに

本研究では，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．

2 種類の段落対での順序推定を行い，まず記事の先頭 2 段落のみの対での順序推定では，提案手法 (0.88) が人間 (0.88) 同程度の性能であった．次に記事内の連接するあらゆる 2 段落対での順序推定では，提案手法 (0.60) が以前の情報との一致数を用いる比較手法 (0.56)) より高い性能であった．

また，文の順序推定を扱った林らの研究結果に基づき，文と段落の順序推定結果を比較した．その結果，記事内先頭 2（文・段落）対の順序を推定することについては，文より段落の方が推定しやすく，記事内のあらゆる連接 2 対の順序を推定することについては，段落より文の方が推定しづらいことがわかった．先頭 2 対の順序を推定する場合では以前の情報がないため，扱う情報が推定する 2 段落のみとなることから，推定情報が多い段落の方が推定しやすくなると思われる．連接 2 対の順序を推定する場合では，段落は各段落内部で話題が完結し前方の文章との関係が小さいため，文より段落の方が推定しづらいと思われる．

第11章　謝辞

本研究を進めるにあたり，終始に渡り研究の進め方や本論文の書き方など，細部に渡る御指導を頂きました，鳥取大学工学部知能情報工学科計算機工学 C 講座の村田真樹教授に心から御礼申し上げます．また，本研究を進めるにあたり，御指導，御助言を頂きました，村上仁一準教授，徳久雅人講師に心から御礼申し上げます．その他様々な場面で御助言を頂きました計算機工学 C 講座研究室の皆様方に感謝の意を表します．

参考文献

[1] Mirella Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 545–552, 2003.

[2] 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, Vol. 2010-NL-199, No. 17, pp. 191–194, 2010.

[3] 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, Vol. 2000-NL-135, No. 11, pp. 55–62, 2000.

[4] 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.

[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, and Mitsuru Ishizuka. A bottom-up approach to sentence ordering for multi-document summarization. In *Information Processing Management*, Vol. 46, pp. 89–109, 2012.

[7] TinySVM: http://chasen.org/~taku/software/tinysvm/.

[8] 村田真樹. 機械学習手法を用いた日本語格解析–教師信号借用型と非借用型，さらには併用型–. 電子情報通信学会技術研究報告, pp. 15–22, 2001.

[9] ChaSen: http://chasen-legacy.sourceforge.jp/.