

概要

文章生成や推敲の問題のひとつに、段落の順序推定が挙げられる。複数の段落からなる文章作成の際に、読み手にわかりやすくなるように段落を適切な順序に並べかえる必要がある。本研究では読み手にわかりやすいように校正された新聞記事を用いて段落の順序推定を行うことで、読み手にわかりやすい段落の順序推定を目指す。

文章を元の順序通りに推定する研究として、伊藤ら [1] の段落の順序推定の研究がある。機械学習を用いて順序が不明な 2 段落に対し、どちらが先かを判定し順序推定を行っている。実験では、記事内最初の 2 段落のみを用いる場合、記事内全ての接続した 2 段落を用いる場合の 2 種類を考慮し順序推定を行い、人間に近い性能を得ている。この研究 [1] では、段落単位での類似性に着目し、順序推定において段落全体の類似性を素性として用いている。しかし、段落間の最後の文と最初の文の類似性である、段落間の隣接する近辺の細かな情報は用いていないという問題がある。段落間の隣接する近辺の細かな情報も用いることで段落の順序推定がしやすくなると考え、そこで、本研究では段落間の隣接する近辺の細かな情報として、段落間の最後の文と最初の文に着目し、伊藤らの素性に拡充していくことで性能向上を目指す。

段落間の隣接する近辺の細かな情報として、段落間の最後の文と最初の文に着目したものを素性として伊藤らの素性に拡充し実験を行った結果、段落の順序を推定する実験において、記事先頭 2 段落のみ用いて順序推定を行った場合、先行研究の素性に新たに素性を拡充した本手法は人手による順序推定と同等の性能であった。記事内の全ての接続 2 段落を用いて順序推定を行った場合、素性を拡充した本手法が先行研究手法に比べてより高い性能であり、人手による順序推定に近づく性能であった。記事内全てから 2 段落を用いて順序推定を行った場合、素性を拡充した本手法は先行研究手法に比べてより高い性能であり、人手による順序推定に近づく性能であった。

また、先行研究手法と本手法及び人手による順序推定の比較により、性能向上がより見られた接続 2 段落の場合において、段落の順序推定にどの素性が有用であるかを検討するために分離平面との距離に基づく素性分析を行い、分析結果を人手で検討した結果、10 個の有用な素性を検討した。

目次

第1章	はじめに	1
第2章	関連研究	3
2.1	順序推定に用いる手法が本研究と異なる関連研究	3
2.2	順序推定の推定対象が本研究と異なる関連研究	3
2.3	用いる文章が本研究と異なる関連研究	4
2.4	素性が本研究と異なる関連研究	5
第3章	問題設定	6
第4章	提案手法	7
4.1	順序が不明な2段落の順序推定の方法	7
4.2	機械学習法	7
4.3	<i>SVM</i> 法	8
4.4	データ作成	10
第5章	提案手法に用いる素性	11
第6章	比較手法	27
第7章	実験	28
7.1	実験条件	28
7.2	入力データの種類	29
7.3	実験結果	32
7.3.1	実験1. 素性拡充前と素性拡充後の提案手法と比較手法による順序推定の比較	32
7.3.2	実験2. 素性拡充前と素性拡充後の提案手法と比較手法と人手推定の比較	32

7.3.3	素性拡充後が正解だった例	33
7.3.4	素性拡充後が不正解だった例	34
7.4	実験の考察	35
第 8 章	SVM 分離平面との距離に基づく素性分析	36
8.1	手法	36
8.1.1	4 で得られた有用な可能性の高い素性を人手により検討する . . .	39
8.2	実験条件	39
8.3	実験結果	40
8.4	素性分析の考察	41
第 9 章	追加実験	42
9.1	段落間の隣接する近辺の細かな情報の類似性による推定法	42
9.1.1	はじめに	42
9.1.2	推定方法	42
9.1.3	実験条件	42
9.1.4	実験結果	43
9.1.5	実験の考察	43
9.2	小説を用いた段落の順序推定	44
9.2.1	はじめに	44
9.2.2	問題設定	44
9.2.3	提案手法	44
9.2.4	提案手法に用いる素性	44
9.2.5	実験条件	44
9.2.6	実験結果	45
9.2.7	実験の考察	45
第 10 章	今後の課題	47
10.1	機械学習に基づく段落の順序推定における素性の拡充	47
10.2	小説を用いた段落の順序推定	47
10.3	上記以外	48
第 11 章	おわりに	49

表 目 次

5.1	素性リスト	11
7.1	各 2 段落対に用いる学習データの 2 段落対の組数 (組)	28
7.2	実験 1 でのテストデータの 2 段落対の組数 (組)	29
7.3	実験 2 でのテストデータの 2 段落対の組数 (組)	29
7.4	素性拡充前と素性拡充後における提案手法と比較手法の正解率	32
7.5	素性拡充前と素性拡充後の提案手法と比較手法と人手推定の正解率	33
8.1	学習データとテストデータの事例数	39
8.2	人手による検討結果 (推定段落 A, B)	41
9.1	素性拡充後における提案手法と隣接情報のみによる推定法と比較手法の 正解率	43
9.2	各データに用いる作品とその詳細	45
9.3	各場合における学習データとテストデータの組数	45
9.4	各場合における提案手法の正解率	45

図 目 次

3.1	問題設定の概念図	6
4.1	機械学習の概念図	7
4.2	マージン最大化	8
5.1	a1 の説明図	12
5.2	a2 の説明図	13
5.3	a3 の説明図	14
5.4	a4 の説明図	15
5.5	a5 の説明図	16
5.6	a6 の説明図	17
5.7	a7 の説明図	17
5.8	a12 の説明図	18
5.9	a20 の説明図	19
5.10	a21 の説明図	20
5.11	1 段落目の最後の文と 2 段落目の最初の文に出現する単語の一致数を求 め方	21
5.12	同順と逆順における単語の一致を求める	22
5.13	段落内における単語番号の図示	23
5.14	境界線の設定	23
5.15	以前の段落の最後の文と推定する各段落の最初の文に出現する単語の一 致数を求める	25
5.16	単語番号の図示と境界線の設定	26
6.1	比較手法 1	27
7.1	入力データ：先頭 2 段落対を用いる場合 (Case1)	30
7.2	入力データ：接続 2 段落対を用いる場合	30

7.3	入力データ：あらゆる2段落対を用いる場合	31
8.1	学習データで用いた個々の素性を素性1個のみの事例を持つテストデータ作成	37
8.2	1の学習データを用いてテストデータを <i>SVM</i> により分類	38
8.3	各事例（素性1個のみ）に対する分離平面との距離	38
8.4	分離平面との距離 d_n が大きい事例ほど有用な可能性の高い素性	39
9.1	推定方法	43

第1章 はじめに

文章生成や推敲の問題のひとつに、段落の順序推定が挙げられる。複数の段落からなる文章作成の際に、読み手にわかりやすくなるように段落を適切な順序に並べかえる必要がある。本研究では校正された新聞記事を用いて段落の順序推定を行うことで、読み手にわかりやすい段落の順序推定を目指す。

文章を元の順序通りに推定する研究として、内元らの語順の推定 [2]、林らの文の順序推定 [3]、伊藤らの段落の順序推定 [1] の研究がある。そのうち、伊藤らの研究では、機械学習を用いて順序が不明な2段落に対し、どちらが先かを判定し順序推定を行っている。実験では、記事内最初の2段落のみを用いる場合、記事内全ての接続した2段落を用いる場合の2種類を考慮し順序推定を行い、人間に近い性能を得た。この研究 [1] では、段落単位での類似性に着目し、順序推定において段落全体の類似性を素性として用いている。しかし、段落間の最後の文と最初の文の類似性である、段落間の隣接する近辺の細かな情報は用いていないという問題がある。段落間の隣接する近辺の細かな情報も用いることで段落の順序推定がしやすくなると考え、そこで、本研究では段落間の隣接する近辺の細かな情報として、段落間の最後の文と最初の文に着目し、伊藤らの素性に拡充していくことで性能向上を目指す。

本論文の特徴を以下に整理する（括弧内は正解率を示す）。

- 出力2分類に対し、学習器に2値分類に秀でた *Support Vector Machine* を用いる。
- 段落間の隣接する近辺の細かな情報が段落の順序推定において有用な特徴と考え、推定段落間の各1文に着目した素性を新たに加えたという新規性がある。
- 段落内の単語の出現位置が段落の順序推定において有用な特徴と考え、素性として新たに加えたという新規性がある。
- 記事の最初の2段落における順序推定では、上記の素性を拡充することにより、拡充する前 (0.85) とほぼ同等の正解率 (0.85) であった。人手との比較を行った

結果、素性の拡充前 (0.88)、後 (0.88) に変わらず人手での順序推定 (0.88) とほぼ正解率であった。

- 記事内の接続する全ての2段落における順序推定では、上記の素性を拡充することにより、拡充する前 (0.60) より高い正解率 (0.62) であった。人手との比較を行った結果、素性の拡充によって拡充前 (0.60) より高い正解率 (0.64) を得た。また、人手による順序推定の正解率 (0.66) に近づいた。
- 記事内全てから2段落を用いる組み合わせにおける順序推定では、上記の素性を拡充することにより、拡充する前 (0.65) より高い正解率 (0.68) であった。人手との比較を行った結果、素性の拡充によって拡充前 (0.72) より高い正解率 (0.75) を得た。また、人手による順序推定の正解率 (0.77) に近づいた。
- より性能向上がみられた接続2段落対の場合を対象に、分離平面との距離に基づく素性分析を行い、段落の順序推定に有用な素性を10個検出した。

2章では関連研究の概要と本研究との違いについて述べる。3章では本研究の問題設定について説明する。4章では本研究が提案する手法や用いる分類器について説明する。5章では機械学習に用いる素性について説明する。6章では提案手法との比較に用いる手法を説明する。6章では段落の順序推定の実験を行い、結果より考察する。7章では素性分析を行い、結果を人手により検討する。8章では2種類の追加実験を行う。9章にまとめを行う。

第2章 関連研究

順序推定に用いる手法，順序推定の対象，用いる文章，素性の4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[4] 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[5] は文章内に一貫性の良くない箇所を推定するために，文章内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．文章の内容が正しく理解されるためには，文章内の質の評価が高くなければならない．そこで，文章の断片中の最後の1文に着目し，先行文脈に対してその文が一貫性かどうかを評価し，具体的に一貫性の悪い箇所を推定することが可能になると考えられる．新聞記事を用いた場合，Entity Grid モデルを断片に対して適用したモデルよりも良い性能であることがわかっている．

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“機械学習”を用いて順序推定を行うという違いがある．機械学習は2値分類に秀でた分類器であり，また素性を簡単に，かつ大量に組み込むことができる．

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[2] は文生成のために，最大エントロピー法を用いて単語の順序を推定する研究を行った．日本語の語順の傾向を整理することが文の解析や生成する際に有効な情報となると考え，文節内外に含まれる様々な情報から語順の傾向を学習するモデルを用いた．正しい語順を京大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

データを京大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。実際には、京大コーパス内の文節の係り受け情報をもとに文節を対象とし順序推定を行っており、語順の傾向を学習するために文節内外の格要素や活用部分の情報を用いることで、有効に働くことがわかっている。

林ら [3] 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文対から元の順序（正例）の文と逆順（負例）の文を作成し、与えられた事例が機械学習によりどの文が先かを判定して文の順序推定を行うというものである。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（新聞記事）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、段落内全ての接続した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種類を考慮し順序推定を行った。さらに、Lapata[4]の手法に基づく確率手法と比較した。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している。

これらの研究では、文節や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

要約文の並び順の決定方法としては、要約対象が複数の文書の場合、文書間の関係も同時に考慮しなければならないため、元の文書中における並び順だけで要約文の順序を決定することはできない。そこで、時間の順序に基づいて並べられた文に対し、岡崎ら [6] 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、時間的な順序で並べられた文の並び順を改善する効果が認められた。

順序の情報は難しいが自然言語の文章を生み出すための重要な課題である。Danushkaら [7] 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機械学習を用いた結果、複数の新聞記事から抽出された文の場合、既存の文の順序推定に対して大幅に改善したと報告している。

これらの研究に対して、本論文では“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問わ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。

2.4 素性が本研究と異なる関連研究

伊藤ら [1] は新聞記事から段落の順序推定のために、多数の素性を用いた機械学習に基づく研究を行った。林ら同様、新聞記事から2段落1組で抜き出し、その2段落対から正例の段落と負例の段落対の問題を作成し、機械学習を用いて、その2段落対の問題が正しい順序かどうかを判定して段落の順序を推定するというものである。実験では、記事内最初の2段落のみを用いる場合、記事内全ての接続した2段落を用いる場合の2種類を考慮し順序推定を行った。また、文の順序推定と段落の順序推定の比較を行い、文と段落の違いを考察し、扱う情報が多いほど推定しやすくなると検討した。

この研究に対して、本論文では、入力の種類として林ら同様 [3] 記事内全てから2段落対の組み合わせから順序を推定する場合も考慮して学習データ、テストデータを作成している。また、段落間の隣接する細かな情報が段落の順序推定に有用であると仮定し、段落間の最後の文と最初の文に着目したものや段落内の単語の出現位置により段落間が似通う傾向から、単語の出現位置を用いて段落間の単語の共起数を求めたものを素性として組み込んでいるという違いがある。

本研究では、伊藤ら [1] の用いた素性を比較として実験を行うこととする。

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いるとする．それより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．確定している段落の次に来る箇所に対し，一方はその箇所に続く段落（正解），もう一方はその箇所に続かない段落（不正解）の2段落を与え，その2段落のうちどちらが先になるかを推定する．順序を推定する際に用いることのできる情報は，推定対象である2段落と推定段落より前の確定している段落とする．

以上の問題設定の概念図を図 3.1 に示す．

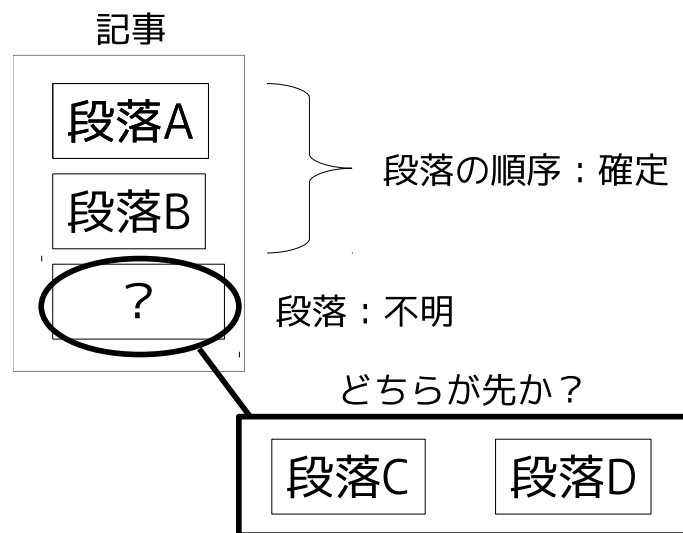


図 3.1: 問題設定の概念図

例として，図 3.1 の場合，段落 A と B は順序が確定しており，その後の段落の順序が不明である．この段落の順序が不明な箇所のうち，確定している段落（段落 B）の次の箇所に対し，段落 C，D のうちどちらが先になるかを推定を行う．推定に用いることのできる情報は推定対象の段落 C，D の2段落と確定している段落 A，B である．

第4章 提案手法

4.1 順序が不明な2段落の順序推定の方法

段落の順序を推定する2段落対が正解タグ（どの段落が先か）で入力された場合，出力が正解となる段落と同じか否かを機械学習で判定する．機械学習には *Support Vector Machine(SVM)*¹を用いる．カーネル関数には2次の多項式カーネルを利用する．

4.2 機械学習法

機械学習法について説明する．図4.1のように，学習器に，扱う事例とその事例に対する出力といった入出力ペアの事例が複数与えられた学習データを学習させておき，事例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．

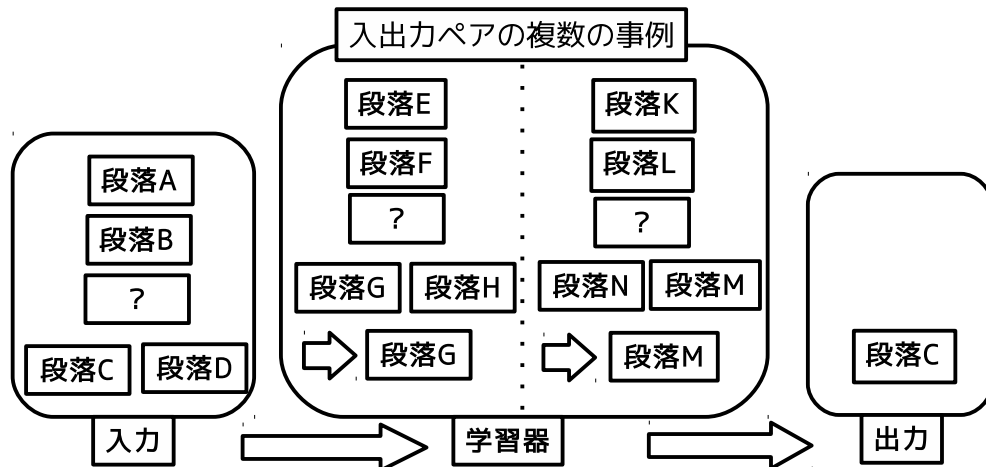


図 4.1: 機械学習の概念図

¹具体的に *SVM* には *TinySVM*[8] 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[9]。このとき、2つの分類が正例と負例からなるものとする、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。

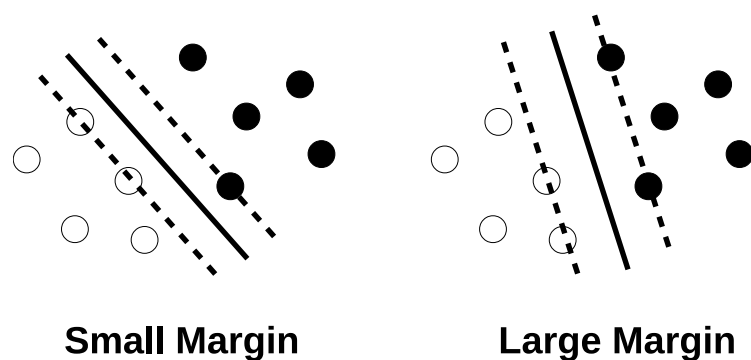


図 4.2: マージン最大化

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

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。

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 $y_i (i = 1, \dots, l, y_i \in \{1, -1\})$ は学習データの文脈と分類先を意味し、関数 sgn は、

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

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。

$$L(\alpha) = \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (4.3)$$

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

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

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を用いる。

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

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2段落対を1組にして抜き出し，その1組の2段落対から元の文章通りの順序（正例）とその逆の順序（負例）の，2つの事例を作成し，それぞれの事例にどの段落が先か（正解）を示す順序タグを付与する．その後，段落内の情報から各素性を求め（素性作成については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タグが付与されるが，*SVM*の出力との正誤判定に用いる．

第5章 提案手法に用いる素性

機械学習で用いられる識別用の特徴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入力を正しく出力できるような素性を学習する。本論文での素性は、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（推定段落 A, B をそれぞれ推定段落 L, R と表記する）。品詞や単語の情報の取得には形態素解析システムの『茶筌』[10]を用いる。本論文で用いる素性を表5に示す（bは本論文で拡充した素性を示す）。素性の説明を以下に示す。

表 5.1: 素性リスト

素性 ID	素性の説明
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	2段落目に出現する名詞の数から1段落目と2段落目に出現する名詞が一致した数で引いた数
a7	素性 a6 の値と推定する2段落を入れ替えた場合の素性 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	2段落目の素性 a2 の前部に出現する名詞の数から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素性 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素性 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	各段落に出現する名詞の数から推定する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素性 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素性 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	各段落の素性 a2 の前部に出現する名詞の数から推定する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素性 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素性 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
a21	1段落目に出現する新規単語数と2段落目に出現する新規単語数との比率

素性 ID	素性の説明
b22	1 段落目の最後の文に出現する単語と 2 段落目の最初の文に出現する単語の一致した数
b23	素性 b22 の値と推定する 2 段落を入れ替えた場合の素性 b22 との比率
b24	段落内の単語の出現位置を用いた場合における 1 段落目に出現する単語と 2 段落目に出現する単語の一致した数
b25	素性 b24 の値と推定する 2 段落を入れ替えた場合の素性 b24 の値との比率
b26	以前の段落の最後の文に出現する単語と推定する各段落の最初の文に出現する単語の一致した数
b27	素性 b26 の値と推定する 2 段落を入れ替えた場合の素性 b26 との比率
b28	段落内の単語の出現位置を用いた場合における以前の段落に出現する単語と推定する各段落に出現する単語の一致した数
b29	素性 b28 の値と推定する 2 段落を入れ替えた場合の素性 b28 の比率

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。段落から茶笥による形態素解析により「品詞：単語」に分け、抽出した情報を素性として用いる。

図 5.1 のような段落が与えられた場合、段落内の文章を分かち書きし、各単語の原形とその品詞が茶笥により出力される。出力結果から下枠内のように「品詞:単語」の形で抽出したものを素性として用いる¹。

段落

「子供を育てる はずの学校には問題が多い」という前提で、教育 現場のユガミをドラマ 化 する。児童の主役はカレーのCMで人気の安達 祐実(子役)。

品詞:単語

名詞:子供 動詞:育てる 名詞:はず 名詞:学校 名詞:問題
 形容詞:多い 名詞:前提 名詞:教育 名詞:現場 名詞:ドラマ 名詞:化
 動詞:する 名詞:児童 名詞:主役 名詞:カレー 名詞:CM 名詞:人気
 名詞:安達 名詞:祐実 名詞:子役

図 5.1: a1 の説明図

¹例. 名詞:子供

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．一般に文中の助詞「は」より前方は新しい内容を表し，助詞「は」より後方は古い内容を示している．この情報を素性として組み込むために以下を行う．まず，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．データ内では，前部に対しを新，後部に対し旧と表現し「(新／旧) 品詞：単語」として素性を組み込んでいる²．

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，全て後部と考える素性とする．

以下の例 5.2 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) 段落を文ごとにわけ，助詞「は」を含む文（文2）は「は」で文を区切り前部と後部でわけ，助詞「は」を含まない文（文1）は後部として用いる．

段落

文1. 「子供を育てる はずの学校には問題が多い」という前提で、教育 現場のユガミをドラマ 化 する。

文2. 児童の主演／は／カレーのCMで人気の安達 祐実(子役)。

区切る

整理

文1.

前部：なし

後部：「子供を育てる はずの学校には問題が多い」という前提で、教育 現場のユガミをドラマ 化 する。

文2.

前部：児童の主演

後部：カレーのCMで人気の安達 祐実(子役)。

図 5.2: a2 の説明図

²例. 旧名詞:子供, 新名詞:児童

a3：段落内文頭に連体詞や接続詞が出現するか否か

段落内の最初の文頭に「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、段落内の最初の文頭に「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。連体詞や接続詞が段落内最初の文頭に出現する場合には1を、出現しない場合には0を素性として用いる。データ内では、「特定品詞：1」または「特定品詞：0」として素性を組み込んでいる。

図 5.3 のように、段落 A 内の最初の文頭に連体詞「この」が出現しているため、段落 A に対し、連体詞が出現するように素性として組み込む³。

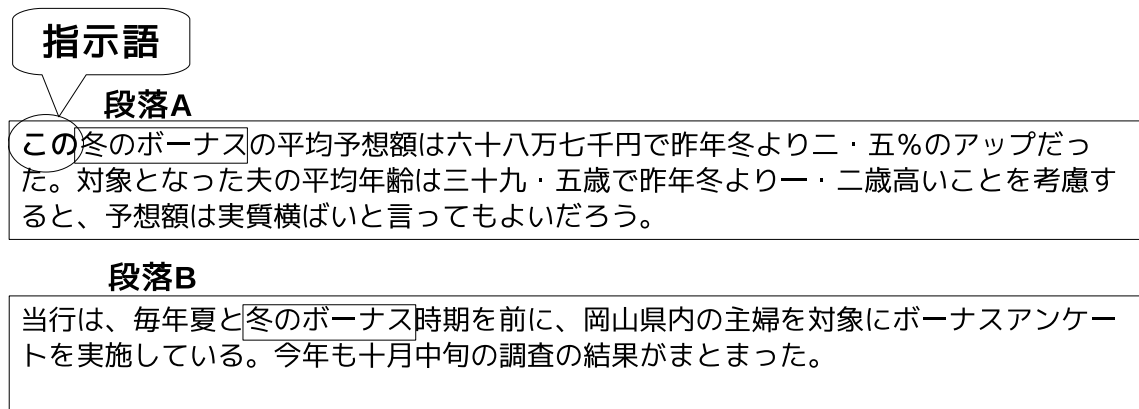


図 5.3: a3 の説明図

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方に記載される傾向がある。この素性はその傾向から名詞「日」に注目し、段落内に上記の単語を含む段落は1を、含まない段落は0を素性として用いる。データ内では、「日付：1」または「日付：0」として素性を組み込んでいる⁴。

図 5.4 のように、段落 A 内に、名詞「日」が出現するため、段落 A に対し、前方となるように素性を組み込む。

³例. 「L 特定品詞:1」段落 A に特定品詞がある, 「R 特定品詞:0」段落 B には特定品詞がない

⁴例. 「L 日付:1」段落 A に日付けがある, 「R 日付:0」段落 B には日付がない

段落A

「男女平等先進国」とされるノルウェーの実態は？という興味深いドキュメンタリー番組が、十八(日)にNHK・BS1で放送される。NHKクリエイティブ制作の「僕の妻は大臣です」。

日付け

段落B

ノルウェー児童家庭省のグレッタ・ベルグ大臣＝写真右＝は今年一月女兒を出産、在職のまま三カ月の育児休暇を取った。日本ではまず考えられないことだ。番組は同大臣にスポットを当てながら、ノルウェーでの夫婦、親子の関係などをリポートする。女性の自立に伴う離婚の増加や子供の養育問題などにも触れる。衛星放送の特徴を生かして、九十分の長尺もの。

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

段落間は似通っているほど、文章の繋がりから読みやすいとされ、また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。推定する2段落の一方に出現する名詞ともう一方に出現する名詞の共起数 X 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 までの区間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、素性 a5 と異なるだけであるため説明を省略する。データ内では、「LR 類似数 $L \circ : X$ (以上)」や「LR 類似数 $L \circ : X-1 \sim X+1$ ($X-1$ から $X+1$ の区間)」として素性を組み込んでいる⁵。

図 5.5 では、各段落それぞれに出現した名詞を示す。段落 A と段落 B の名詞の共起数 (7) を求め、その値を場合わけし素性として組み込む。

⁵例. 「LR 類似数 $L \circ : 7$ 」段落 AB 間での一致数が 7 個以上、「LR 類似数 $L \circ : 6 \sim 8$ 」段落 AB 間での一致数が 6 から 8 の間にある

段落(A)

(名詞情報：)
運輸省，空港，街路，灯，
定期，点検，長年，年，
実施，経費，年間，こと，
会計検査院，十分，改善，
措置，要求，管理，
建設省，無駄遣い，浮き彫り

段落(B)

指摘，空港，道路，歩道，
街路，灯，点検，今年，初め，
全国，主要，実施，結果，
管理，航空局，事務所，業者，
委託，年

AとBの名詞の一致した数X：7つ

$0 \leq \dots \leq 6 \leq X$ かつ $6 \leq X < 8$

図 5.5: a5 の説明図

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

用いられない単語の数が推定の鍵になると考え素性として組み込む。素性 a6 は 2 段落目に出現する名詞のうち，1 段落目の名詞と一致しなかった名詞の数 X 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。データ内では，「LR 類似数 $L \times : X$ 」や「LR 類似数 $L \times : X-1 \sim X+1$ 」として素性を組み込んでいる（「LR 類似数 $L \times : X$ 」の LR は推定する 2 段落を示し， $L \times$ は段落 B ではなく段落 A にのみ出現する単語について示し，「：」以降に求めた値を示す）。

図 5.6 のように，段落 B の名詞の数 (19) から段落 A と段落 B 内の名詞の共起数 (7) を引いた数 (12) を場合わけし素性として組み込む⁶。

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

順序を推定する 2 段落の段落 A，B において，「 $A \rightarrow B$ 」という順序の 2 段落に対し A ではなく B にのみ出現する名詞の数 X と，逆の「 $B \rightarrow A$ 」という順序の 2 段落に対し B ではなく A にのみ出現する名詞の数 Y を求める。（Z=） $X-Y$ を行い，その値が 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 a10，素性 a11 も同様であるた

⁶例. 「LR 類似数 $L \times : 10$ 」段落 AB 間で A にのみ出現する単語数 10 個以上，「LR 類似数 $L \times : 8 \sim$ 」段落 AB 間で A にのみ出現する単語 8 以上

段落(A)

(名詞情報：)
運輸省，空港，街路，灯，
定期，点検，長年，年，
実施，経費，年間，こと，
会計検査院，十分，改善，
措置，要求，管理，
建設省，無駄遣い，浮き彫り

段落(B)

指摘，空港，道路，歩道，
街路，灯，点検，今年，初め，
全国，主要，実施，結果，
管理，航空局，事務所，業者，
委託，年

AとBの名詞の一致した数：7つ

Bの名詞の数：19つ

⇒一致しなかった数Y：12つ

$0 \leq \dots \leq 10 \leq Y$ かつ $8 \leq Y$

図 5.6: a6 の説明図

め，省略する．データ内では，「LR 類似度 $L \times : 0$ 」または「LR 類似度 $L \times : X \sim X+2$ 」として素性を組み込んでいる⁷．

図 5.7 のように，「 $A \rightarrow B$ 」における B 内で一致しなかった数 (12) と「 $B \rightarrow A$ 」における A 内で一致しなかった数 (14) を求め，求められた 2 つの差 (2) を場合わけし素性として組み込む．

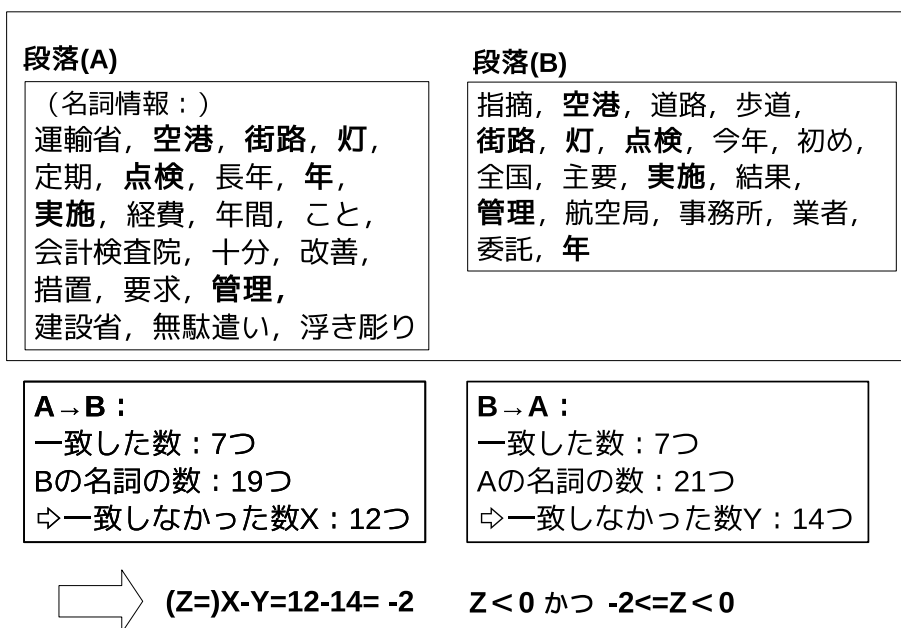


図 5.7: a7 の説明図

⁷例. 「LR 類似度 $L \times : 0$ 」 A にのみ出現する単語数から B にのみ出現する単語数を引いた数が 0 以上

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内容の理解がしやすいため、文章の順序としては良い。このことから推定する2段落と存在する記事内以前全ての段落との名詞の共起した数をそれぞれ求め、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までの範囲で場合わけしたものを素性とする。素性a12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。データ内では、「P類似度 $L \circ : X$ 」または「P類似数 $L \circ : X \sim X+2$ 」として素性を組み込んでいる（ここで、Pは以前の段落を表す）⁸。

図5.8のように、推定する2段落A, Bより以前の確定している段落Pが与えられたとする。「P → A」間、「P → B」間の名詞の共起数をそれぞれ求める(4), (1)。求められた2つの値を場合わけし素性として組み込む。

以前の段落P

(名詞情報:)
食糧庁, 凶作, 緊急, 輸入,
外国, 産米, 家畜, 飼料,
売却, 正式, 対象, 中国,
中心, 程度, 売買, 保管

段落A

緊急, 輸入, 米, 現在, 政府,
在庫, 飼料, 売却, 着色, 異物,
混入, 品質

段落B

主食, 大量, 飼料, 処理, 政府,
在庫, 過剰, 以来

段落P-A間: 名詞の共起数4つ } 共起数: 段落P-A間 > 段落P-B間
段落P-B間: 名詞の共起数1つ } 段落Aが先になるような素性を付与

図 5.8: a12 の説明図

⁸例. 「P類似数 $L \circ : 4$ 」PA間の単語の一致数が4個以上, 「P類似数 $L \circ : 4 \sim 6$ 」PA間の単語の一致数が4から6の間にある

a20：1 段落目と2 段落目に出現する新規単語の数の差

話が変わると、今まで出てこなかった単語が初めて出現することが多い。このことから、初めて出現する単語（新規単語）の数を求め、推定する2段落のうち、より新規単語の数が多い段落を後の段落になりやすいように素性として用いる。新規単語の条件は以前の段落には出現していないもののみとする。

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。データ内では、「LR 新規：0 超過」または「LR 新規:0 未満」として素性を組み込んでいる⁹。

図5.9のように、以前の段落と推定段落A、Bが与えられたとする。以前の段落と各段落との単語の共起数をそれぞれ求め(4)、(3)、その値から以前の段落に対し各段落にのみ出現する単語の数を求める(14)、(26)。求められた2つの値の差によって得られた値(12)を場合わけし素性として組み込む。

以前の段落 (出現した単語：) 企業、経常、利益、大幅、 見込む、いる、売り上げ、 伸び悩む、注目、個別、品目、 横ばい、リストラ、徹底的、 やる、効果、しかし、無限、 できる、価格、破壊、 まだ、始まる、体験、 なる、今後、どんな、影響、 出る、する、	段落A 設備、投資、中、回復、する、こと、改めて、 注目、早い、前半、みる、いる、雇用、 あまり、改善、リストラ、続く、ため 段落B 短観、景気、回復、裏付ける、基盤、強い、 ため、今後、機動、政策、運営、必要、 来年度、当初、予算、今、対策、やる、 過ぎる、いる、急激、落ち込む、かねる、 公共、投資、追加、タイミング、良く、ある
--	---

A：以前に出現する単語4つ、全体単語数18つ → 新規単語数X14つ

B：以前に出現する単語3つ、全体単語数29つ → 新規単語数Y26つ



$$X-Y=14-26=-12<0$$

図 5.9: a20 の説明図

⁹例. LR 新規:0 超過, LR 新規:0 未満

a21：1 段落目と2 段落目に出現する新規単語の比率

新規単語の比率は素性 a20 同様新規単語の数を段落 A, B に対してそれぞれ求め、求めた 2 つの値の比率を素性とする。ここで用いる品詞も素性 a20 同様のものとする。データ内では、「LR 新規比率:L」または「LR 新規比率:R」として素性を組み込んでいる¹⁰。

図 5.10 のように、図 5.9 で求めた新規単語数それぞれを用いて、A の B に対する比率 (0.54) と B の A に対する比率 (1.86) を求め、求めた 2 つの値により、A の B に対する比率の方が大きいため、B の方が後方に来る、すなわち A が前方となるような素性を組み込む。

以前の段落P

(出現した単語:)

企業, 経常, 利益, 大幅,
見込む, いる, 売り上げ,
伸び悩む, 中奥, 個別, 品目,
横ばい, リストラ, 徹底的,
やる, 効果, しかし, 無限,
できる, 価格, 破壊, まだ,
始まる, 体験, なる, 今後,
どんな, 影響, 出る, する

段落A

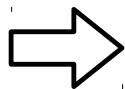
設備, 投資, 中, 回復, する, こと, 改めて,
注目, 早い, 前半, みる, いる, 雇用, あまり,
改善, リストラ, 続く, ため

段落B

短観, 景気, 回復, 裏付ける, 基盤, 強い,
ため, 今後, 機動, 政策, 運営, 必要, 来年度,
当初, 予算, 今, 対策, やる, すぎる, いる,
急激, 落ち込む, かねる, 公共, 投資, 追加,
タイミング, 良く, ある

段落P-A間：単語の共起数4つ，段落A全体単語数18つ → 新規単語数14つ

段落P-B間：単語の共起数3つ，段落B全体単語数29つ → 新規単語数26つ



比率A/B : $14/26=0.54$

比率B/A : $26/14=1.86$

比率A/B < 比率B/A

段落Aが先になるような素性を付与

図 5.10: a21 の説明図

¹⁰例. LR 新規比率:L

b22：1段落目の最後の文に出現する単語と2段落目の最初の文に出現する単語の一致した数

この素性は、1段落目の最後の1文に出現する単語と2段落目の最初の1文に出現する単語との共起した数 X を用いる（図 5.11）。

また、求められた値： X がどの辺りに属するか、区間を設ける。 X が、0 以上、1 以上、2 以上、3 以上を 1 ずつ増加し最大 10 まで、0 以上 1 未満、1 以上 2 未満、2 以上 3 未満を 1 ずつ増加し最大 10 までの範囲で場合わけとする。データ内では、「一文一致数： X 」または「一文区間： $X \sim X+1$ 」として素性を組み込んでいる¹¹。

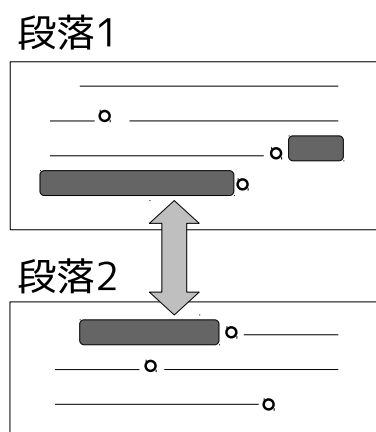


図 5.11: 1段落目の最後の文と2段落目の最初の文に出現する単語の一致数を求め方

b23：素性 b22 の値と推定する2段落を入れ替えた場合の素性 b22 との比率

1段落目の最後の1文に出現する単語と2段落目の最初の1文に出現する単語の一致した数 A を求め、さらに推定する2段落を入れ替えた場合の前述の単語の一致した数 B を求める。求めた値から A の B に対する比率 X を算出し素性として用いる（図 5.12）。

また、求められた値： X がどの辺りに属するか、同様に区間を設ける。 X が、0 と同値、0 以上、0.5 以上、1 以上、1.5 以上を 0.5 ずつ増加し最大 5 まで、0 以上 0.5 未満、0.5 以上 1 未満、1 以上 1.5 未満を 0.5 ずつ増加し最大 5 までの範囲で場合わけとする。

¹¹例. 一文一致数: 4, 一文区間: 4 ~ 5

データ内では、「一文比率一致数：X」または「一文比率区間：X～X+0.5」として素性を組み込んでいる¹²。

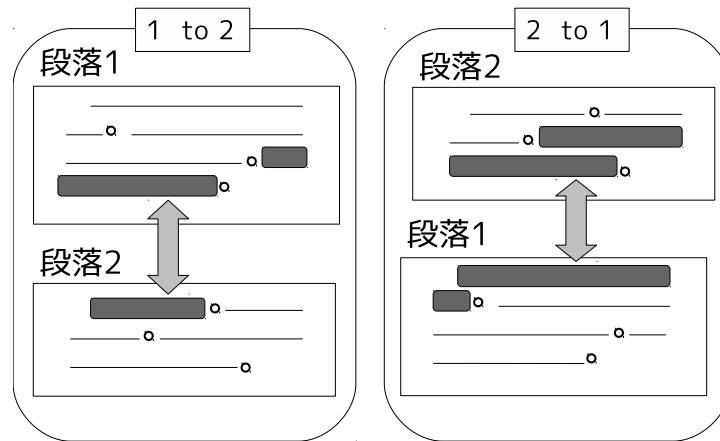


図 5.12: 同順と逆順における単語の一致を求める

b24：段落内の単語の出現位置を用いた場合における 1 段落目に出現する単語と 2 段落目に出現する単語の一致した数

1 段落目に出現する単語と 2 段落目に出現する単語の一致した数を素性と用いており、またこの素性は出現した単語の位置が組み込まれている。まず 1 段落目と 2 段落目との間に境界線を敷く。その境界線を 1 とし境界線から遠ざかるほど値が低くなるよう 1 段落目の最初と 2 段落目の最後を 0 と設定する (図 5.13, 図 5.14)。各段落内の単語に境界線からどの位置に存在するかを表す値 (以下、距離値とする) を算出し、単語に付与する。1 段落目に出現する単語 A と 2 段落目に出現する単語 B が単語一致した場合、“単語 A に付与された距離値 * 単語 B に付与された距離値”により算出された値をスコアとし、単語一致するごとに加算する。総加算のスコア X を素性として用いる。また、1, 2 段落内で一致する単語の複数の共起が見られた場合、その単語の距離値の総和を共起の組数で割った値とする。

また、求められた値：X がどの辺に属するか、同様に区間を設ける。X が、0 以上、0.5 以上、1 以上、1.5 以上を 0.5 ずつ増加し最大 10 まで、0 以上 0.5 未満、0.5 以上 1 未満、1 以上 1.5 未満を 0.5 ずつ増加し最大 10 までの範囲で場合わけとする。データ内では、「一致数：X」または「区間：X～X+0.5」として素性を組み込んでいる¹³。

¹²例. 一文比率一致数: 4、5、一文比率区間: 4、5～5

¹³例. 一致数: 8、区間: 8～8、5

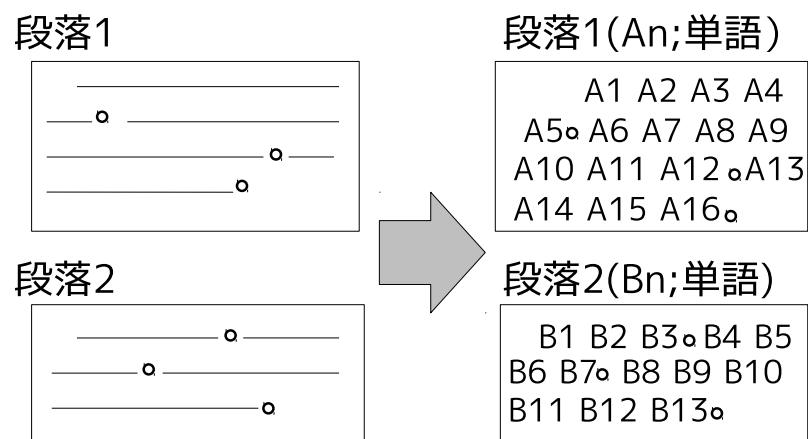


図 5.13: 段落内における単語番号の図示

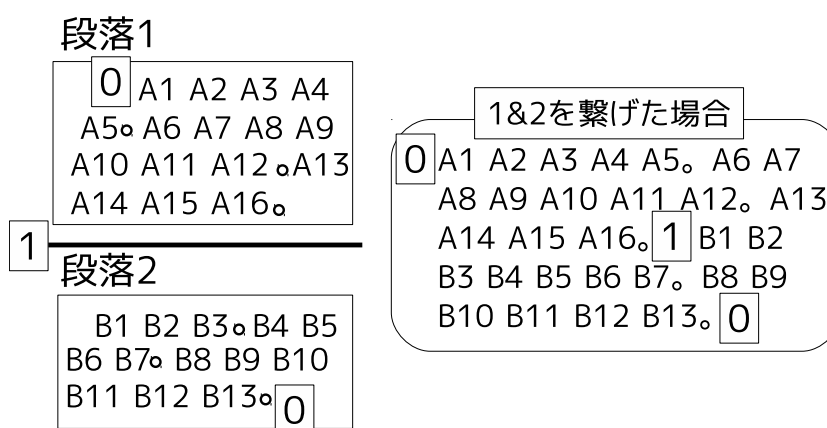


図 5.14: 境界線の設定

b25：素性 b24 の値と推定する 2 段落を入れ替えた場合の素性 b24 の値との比率

b24 同様この素性は出現した単語の位置が組み込まれている。1 段落目に出現する単語と 2 段落目に出現する単語の一致した数を求め、さらに推定する 2 段落を入れ替えた場合の前述の単語の一致した数を求める。求めた値から比率 X を算出し素性として用いる。

また、求められた値： X がどの辺に属するか、同様に区間を設ける。 X が、0 と同値、0 以上、0.5 以上、1 以上、1.5 以上を 0.5 ずつ増加し最大 5 まで、0 以上 0.5 未満、0.5 以上 1 未満、1 以上 1.5 未満を 0.5 ずつ増加し最大 5 までの範囲で場合わけとする。データ内では、「比率一致数： X 」または「比率区間： $X \sim X+0.5$ 」として素性を組み込んでいる¹⁴。

b26：以前の段落の最後の文に出現する単語と推定する各段落の最初の文に出現する単語の一致した数

推定する 2 段落以前の段落の最後の 1 文に出現する単語と推定する各段落の最初の 1 文に出現する単語の一致した数 X を素性として用いている (図 5.15)。

また、求められた値： X がどの辺りに属するか、区間を設ける。 X が、0 以上、1 以上、2 以上、3 以上を 1 ずつ増加し最大 10 まで、0 以上 1 未満、1 以上 2 未満、2 以上 3 未満を 1 ずつ増加し最大 10 までの範囲で場合わけとする。データ内では、「P 一文一致数： X 」または「P 一文区間： $X \sim X+1$ 」として素性を組み込んでいる¹⁵。

b27：素性 b26 の値の比率

推定する 2 段落 A、B の以前の段落の最後の 1 文に出現する単語と推定する各段落の最初の 1 文に出現する単語の一致した数から P-A の P-B に対する比率を算出し素性として用いる。

また、求められた値： X がどの辺に属するか、同様に区間を設ける。 X が、0 と同値、0 以上、0.5 以上、1 以上、1.5 以上を 0.5 ずつ増加し最大 5 まで、0 以上 0.5 未満、

¹⁴例. 比率一致数: 3, 比率区間: 3 ~ 3、5

¹⁵例. P 一文一致数: 9, P 一文区間: 9 ~ 10

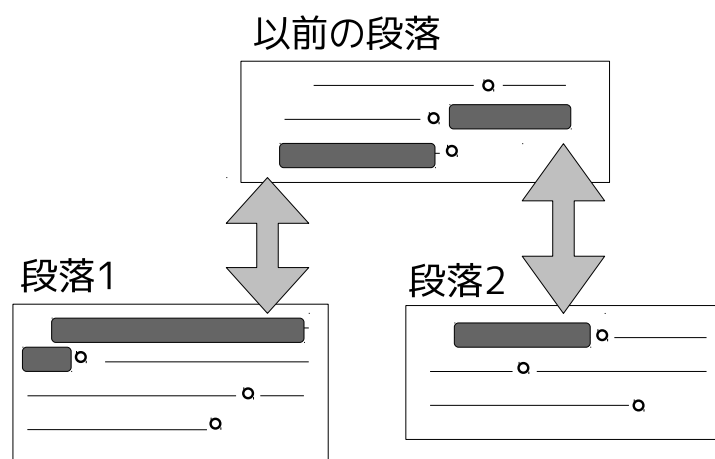


図 5.15: 以前の段落の最後の文と推定する各段落の最初の文に出現する単語の一致数を求める

0.5 以上 1 未満, 1 以上 1.5 未満を 0.5 ずつ増加し最大 5 までの範囲で場合わけとする. データ内では, 「P 一文比率一致数: X」または「P 一文比率区間: X~X+0.5」として素性を組み込んでいる¹⁶.

b28: 以前の段落に出現する単語と推定する各段落に出現する単語の一致した数

推定する 2 段落以前の段落に出現する単語と推定する各段落に出現する単語の一致した数を素性としている. この素性は素性 b24 同様, 出現した単語の位置情報が組み込まれている. まず 1 段落目と 2 段落目との間に境界線を敷く. その境界線を 1 とし境界線から遠ざかるほど値が低くなるよう 1 段落目の最初と 2 段落目の最後を 0 と設定する (図 5.16). 各段落内の各単語における距離値を算出し, 単語に付与する. 1 段落目に出現する単語 A と 2 段落目に出現する単語 B が単語一致した場合, “単語 A に付与された距離値 * 単語 B に付与された距離値” により算出された値をスコアとし, 単語一致するごとに加算する. 総加算のスコア X を素性として用いる. また, 1, 2 段落内で一致する単語の複数の共起が見られた場合, その単語の距離値の総和を共起の組数で割った値とする.

また, 求められた値: X がどの辺に属するか, 同様に区間を設ける. X が, 0 以上, 0.5 以上, 1 以上, 1.5 以上を 0.5 ずつ増加し最大 10 まで, 0 以上 0.5 未満, 0.5 以上 1 未

¹⁶例. P 一文比率一致数: 3, P 一文比率区間: 3~3, 5

満, 1 以上 1.5 未満を 0.5 ずつ増加し最大 10 までの範囲で場合わけとする. データ内では, 「P 一致数: X」 または 「P 区間: X~X+0.5」 として素性を組み込んでいる¹⁷.

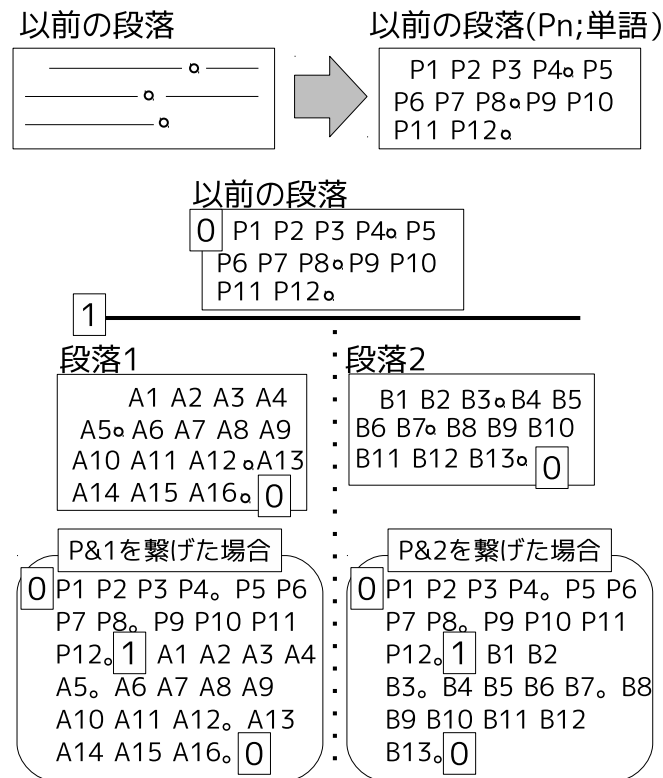


図 5.16: 単語番号の図示と境界線の設定

b29：素性b28の値の比率

素性 b28 同様, この素性は出現した単語の位置が組み込まれている. 推定する 2 段落 A, B 以前の段落 P に出現する単語と推定する各段落に出現する単語の一致した数から, P-A の P-B に対する比率 X を算出し素性として用いている.

また, 求められた値: X がどの辺に属するか, 同様に区間を設ける. X が, 0 と同値, 0 以上, 0.5 以上, 1 以上, 1.5 以上を 0.5 ずつ増加し最大 5 まで, 0 以上 0.5 未満, 0.5 以上 1 未満, 1 以上 1.5 未満を 0.5 ずつ増加し最大 5 までの範囲で場合わけとする. データ内では, 「P 比率一致数: X」 または 「P 比率区間: X~X+0.5」 として素性を組み込んでいる¹⁸.

¹⁷例. P 一致数: 4, P 区間: 4~4、5

¹⁸例. P 比率一致数: 5, P 比率区間: 5~5、5

第6章 比較手法

提案手法の有用性確認のために，以下の手法を用いて比較する．

連接する2段落は内容が似通うことから，推定する2段落以前の段落に出現する名詞と推定する各段落に出現する名詞の共起数をそれぞれ求め，単語の共起数がより大きい方の段落を先とする手法である．

図6.1のように，推定する2段落A，Bと推定段落以前の段落Pにおいて，PA間とPB間の名詞の共起数X，Yをそれぞれ求める．Xが3，Yが2となる場合，段落PA間の内容が似通っていると判断し，段落Aが先とする．

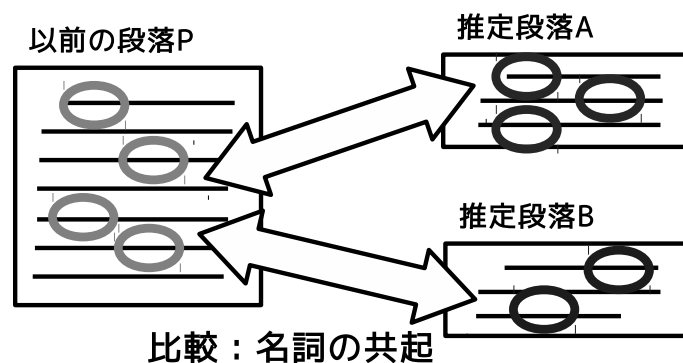


図 6.1: 比較手法 1

第7章 実験

7.1 実験条件

実験で順序推定を行う際に用いる2段落対の組には3種類の場合を考慮して学習データとテストデータを作成する.

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 Case1 と表記)
2. 記事内のあらゆる接続する2段落の対の順序を推定する場合 (以下 Case2 と表記)
3. 記事内全てから2段落を用いる組み合わせの対の順序を推定する場合 (以下 Case3 と表記)

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: 各2段落対に用いる学習データの2段落対の組数 (組)

	Case1	Case2	Case3
学習データ	3,124	15,020	80,248

本実験では, 以下3種類の手法を用い,

1. 先行研究 [1] のみの素性を用いる提案手法 (以下素性拡充前と表記)
 2. 先行研究の素性に本研究で作成した素性を拡充した提案手法 (以下素性拡充後と表記)
 3. 被験者2名の順序推定の平均 (以下人手推定と表記)
- 2種類の比較実験 (実験 1, 実験 2) において各2段落対での順序推定を各々行う.

実験 1. 素性拡充前と素性拡充後の提案手法と比較手法による順序推定の比較

実験 2. 素性拡充前と素性拡充後の提案手法と比較手法と人手推定の比較

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、Case1 の順序推定を行う場合は毎日新聞 1993 年 6 月の 1 ヶ月分、Case2 の順序推定を行う場合は毎日新聞 1993 年 7 月の 1 ヶ月分、Case3 の順序推定を行う場合は毎日新聞 1993 年 8 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	Case1	Case2	Case3
テストデータ	412	3,146	13,040

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	Case1	Case2	Case3
テストデータ	50	50	50

7.2 入力データの種類

1. 記事内の最初の 2 段落のみを用いる場合

入力データが先頭 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（正例）とその逆順（負例）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正解の順序タグをそれぞれの事例に付与し、テストデータもまた学習データ同様、各事例に順序タグをそれぞれの事例に付与するが、テストデータの場合は機械学習からの出力による推定結果との正誤判定の際に用いる。

記事内の先頭 2 段落であり、推定する 2 段落対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 から素性 a21, 素性 b26 から素性 b29）を用いることができない。

図 7.1 を例に挙げると、段落 A, B, C の順序からなる 1 つの記事から、抽出される先頭 2 段落対は記事内の初めの 2 段落対であるため、(A, B) になる。抽出された 2 段

¹テストデータ：ランダムに 2 段落 1 組を 50 組抽出

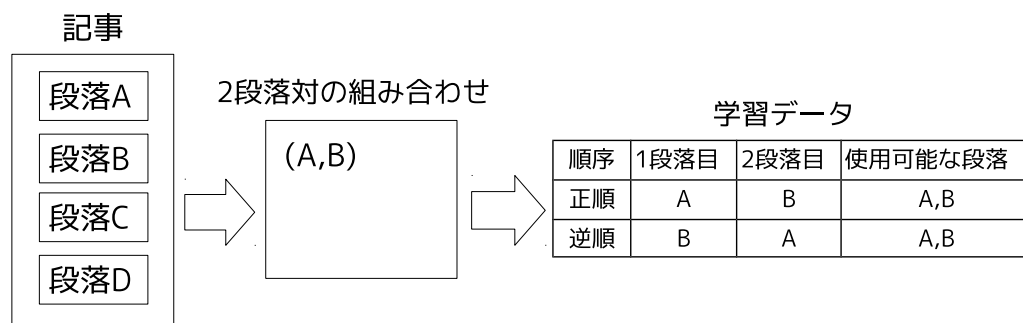


図 7.1: 入力データ：先頭 2 段落対を用いる場合 (Case1)

落対（：A，B）から正順（：A→B）と逆順（：B→A）をそれぞれ入出力データとして学習データ，テストデータを作成する．また，先頭 2 段落対を用いる場合なので，順序推定に用いることができる情報は推定段落（A，B）となる．

2. 記事内のあらゆる接続する 2 段落を用いる場合 (Case2)

入力データが接続 2 段落対の場合では，記事内のあらゆる接続する 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，作成に用いた記事での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タグをそれぞれの事例に付与し，テストデータもまた学習データ同様，各事例に順序タグをそれぞれの事例に付与するが，テストデータの場合は機械学習からの出力による推定結果との正誤判定の際に用いる．

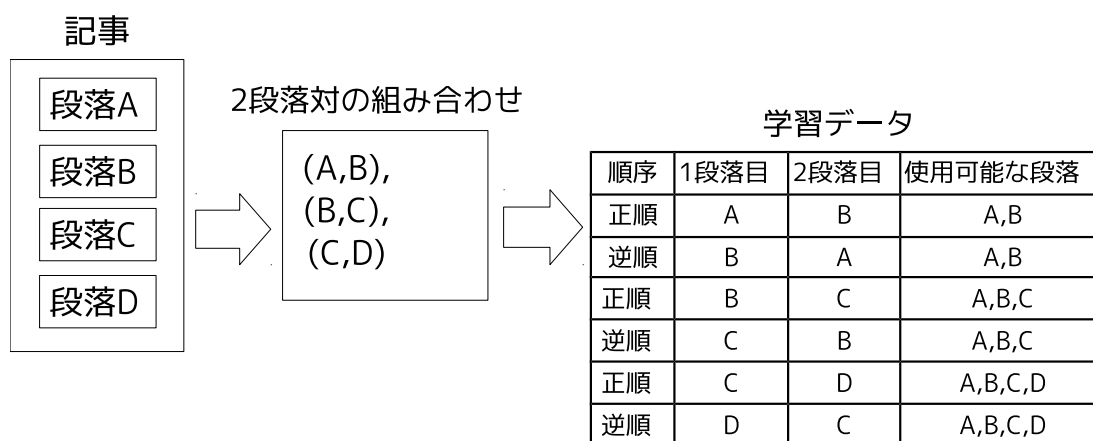


図 7.2: 入力データ：接続 2 段落対を用いる場合

図 7.2 を例に挙げると，段落 A，B，C の順序からなる 1 つの記事から，抽出される

接続2段落対は記事内のあらゆる接続する2段落対であるため、(A, B), (B, C) となる。抽出された2段落対 (:A, B) から正順 (:A → B) と逆順 (:B → A) をそれぞれ入出力データとして学習データ、テストデータを作成する ((B, C) からは正順 (:B → C) と逆順 (:C → B) を作成)。また、接続2段落対を用いる場合なので、順序推定に用いることのできる情報は (A, B) の場合は推定段落 (A, B) となり、(B, C) の場合は推定段落 (B, C) と以前の段落 A の (A, B, C) となる。

3. 記事内全てから2段落を用いる組み合わせの対の順序を推定する場合 (Case3)

入力データが記事内全てから2の段落対の場合では、記事内全ての2段落の組み合わせの対を用いて、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タグをそれぞれの事例に付与し、テストデータもまた学習データ同様、各事例に順序タグをそれぞれの事例に付与するが、テストデータの場合は機械学習からの出力による推定結果との正誤判定の際に用いる。

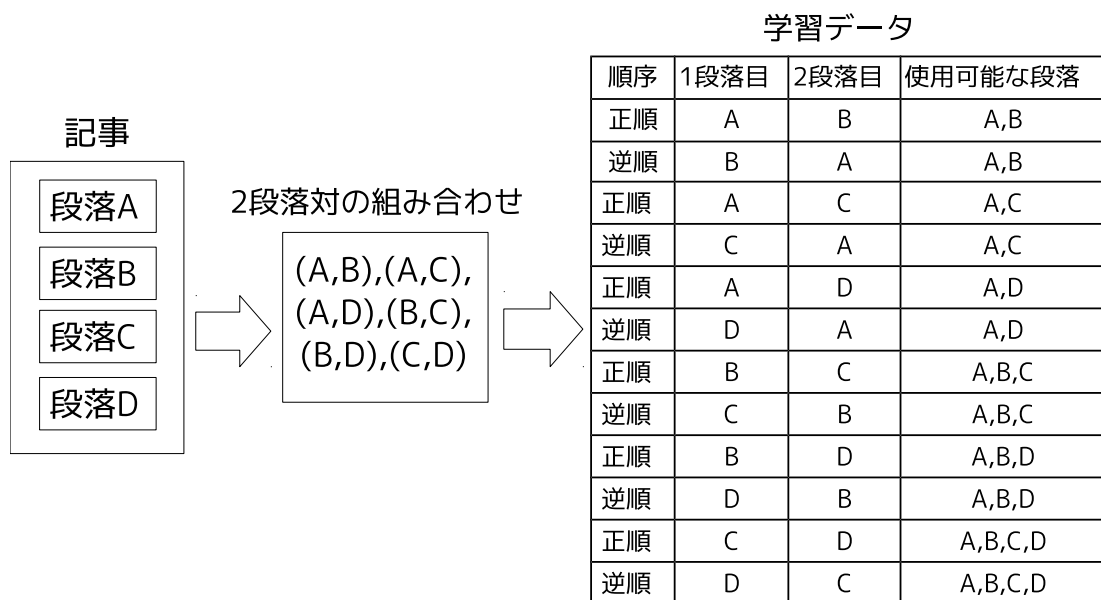


図 7.3: 入力データ：あらゆる2段落対を用いる場合

図 7.3 を例に挙げると、段落 A, B, C, D の順序からなる1つの記事から、抽出される2段落対は記事内全ての2段落対であるため、(A, B), (A, C), (B, C), (B,

D), (C, D) となる。抽出された2段落対 (:A, B) から正順 (:A → B) と逆順 (:B → A) をそれぞれ入出力データとして学習データ, テストデータを作成する他の2段落対も同様)。また, あらゆる2段落対を用いる場合なので, 順序推定に用いることのできる情報は (A, B) の場合は推定段落 (A, B) となり, (B, C) の場合は推定段落 (B, C) と以前の段落 A の (A, B, C) となり, (B, D) の場合は推定段落 (B, D) と推定段落どちらか一方が出現するまでにある以前の段落 A の (A, B, D) となる。

7.3 実験結果

7.3.1 実験 1. 素性拡充前と素性拡充後の提案手法と比較手法による順序推定の比較

素性の拡充前と拡充後の提案手法と比較手法の順序推定の正解率を表 7.4 に示す。素性拡充前と拡充後に対して両側検定の有意差水準 5% で検定を行った結果, Case1 において有意差なし, Case2 において有意差あり, Case3 において有意差ありであった (表 7.4 内の*は有意差ありを示す)。

表 7.4: 素性拡充前と素性拡充後における提案手法と比較手法の正解率

入力 段落対	提案手法		比較 手法
	素性拡充前	素性拡充後	
Case1	0.85	0.85	-
Case2	0.60	0.62*	0.53
Case3	0.65	0.68*	0.62

表 7.4 から, 素性を拡充した場合の提案手法が Case2,3 において, 素性を拡充する前より高い正解率であることがわかる。また, 有意差検定により, Case1 では, 素性の拡充は有意に性能を向上させることには役立たなかった。Case2,3 では, 素性の拡充が役立つことが確認できた。

7.3.2 実験 2. 素性拡充前と素性拡充後の提案手法と比較手法と人手推定の比較

人による順序推定も加え, 比較を行う。素性拡充前と素性拡充後の提案手法と人手推定の正解率を表 7.5 に示す。素性拡充前と拡充後に対して両側検定の有意差水準 5% で

検定を行った結果、Case1において有意差なし、Case2において有意差なし、Case3において有意差ありであった（表7.5内の*は有意差ありを示す）。

表 7.5: 素性拡充前と素性拡充後の提案手法と比較手法と人手推定の正解率

入力 段落対	提案手法		比較 手法	人手推定		
	素性拡充前	素性拡充後		A	B	平均
Case1	0.88	0.88	-	0.92	0.84	0.88
Case2	0.60	0.64	0.56	0.68	0.64	0.66
Case3	0.72	0.75*	0.56	0.84	0.70	0.77

表7.5から、素性を拡充した場合の提案手法がCase2,3の場合に、素性を拡充するより高い正解率であり、人手推定に近づく正解率であった。

7.3.3 素性拡充後が正解だった例

本節では、素性拡充前が不正解で、素性拡充後が正解だった2例を記載する。

（例1）正解：段落A

段落A：この作家は、歴史家でもあるウラジーミル・ゴニク氏。**資料を集め、関係者の取材を続けたうえ、自ら「第二の地下鉄」を歩いて調査した。**

段落B：この地下鉄は、**第二次世界大戦後、現在の地下鉄より深いところに掘られた。**長さは数十キロとされるが、正確な距離は不明。クレムリンからスターリンの別荘があったクンツェフ地区に通じる視線のほか、国家保安委員会（KGB）、外務省など国家の重要施設からトンネルを通り地下鉄に乗れる仕組みになっている。

例1では、段落A→Bの順序の場合に段落間の各文において単語「地下鉄」の共起がみられるため、段落A→B間の内容は似通っていると判断し、段落Aが先にくると推定したと思われる。

(例2) 正解：段落 A

以前の段落 P：リクルートサーチがまとめた今年の大卒求人倍率調査（対象は来春卒業者）によると、大卒男子の求人倍率は二・二二倍で、一九八四年の調査開始以来最低となった。**全国民間企業の求人総数は昨年より十万人少ない五十二・八万人。**

段落 A：**求人数を企業規模別に見ると、従業員千人未満の企業で昨年の一三%減なのに比べ、千人以上の企業では二二%減。**

段落 B：**また、文化系の求人倍率は二・〇一倍、理科系は二・五一倍。**落ち込み幅は理科系の方が大きく、証券など非製造業での理科系採用減が響いたようだ。

例2では、例1同様に以前の段落 P → A、段落 P → B の順序の場合に段落間の各文において共に単語の共起がみられるため、それぞれの順序での各1文に対する単語の共起数を求めた結果、P → A では5つ（求人、企業、昨年）、P → B では1つ（求人）の単語の共起がみられた。この結果より、以前の段落 P と段落 A の方がより段落間の内容が似通っていると判断し、段落 A が先にくると推定したと思われる。

7.3.4 素性拡充後が不正解だった例

本節では、素性拡充前が正解で、素性拡充後が不正解だった2例を記載する。

(例3) 正解：段落 B

段落 A：昨年のコメ生産量は全体で約千十九万トン。うち政府米百五十七万トン、自主流通米四百五十三万トンの計六百十万トンが政府管理米。残る四百九万トンは、一部は農家が自分で消費し、大半がヤミ米に流れた。**政府米は政府管理米の二六%にとどまった。**

段落 B：**政府米に比べ、六十キロ当たり五千円以上も農家の手取りが高い自主流通米が、全体の七割前後を占めている現在、生産者米価は価格を下支えする機能すら失いつつあるのが実態。**

例3では、段落 A → B の順序の場合に段落間の各文において、単語「政府」、「米」の共起がみられるため、段落 A → B 間の内容は似通っていると判断し、段落 A が先にくると推定したが誤りであった。

(例4) 正解の順序：段落 A

段落 A：この日は全国から参加した男子七十九人が制限時間 8 分と 10 分の A・B ニコースに挑戦した=写真。壁は高さ二〇メートル。完登はただ一人だった。

段落 B：最高気温三一・八度の猛暑でバテたのか、途中で力尽き中ぶらりんとなる選手が続出。出場者の中には「やっぱり、ひんやりした北アルプスの頂の方がいい」の声も。一日も行われる。

例4では、段落 B → A の順序の場合に段落間の各文において、単語「日」の共起がみられるため、段落 A → B 間の内容が似通っていると判断し、段落 B が先にくると推定したが誤りであった。

7.4 実験の考察

表 7.4, 表 7.5 より, Case1 の場合に素性を拡充する前と拡充した後の提案手法はほぼ同等の正解率であることがわかる. このことは Case1 の場合において, 拡充する素性より今までの素性の方が有用な素性であったことがわかる. また, Case2,3 の場合にどちらも素性を拡充した本手法の性能が高いことがわかる. このことは Case2,3 の場合において, 段落間の隣接する近辺の細かな情報を着目し推定対象の 2 段落間の結びつきの特徴を考慮した素性 (b22,b23,b26,b27) や以前の段落と推定対象の各段落間の結びつきの特徴を考慮した素性 (b24,b25,b28,b29) を用いたことにより, 正解率が素性を拡充する前に比べて上昇したのではないかと推測する.

そこで, 素性分析により本手法で用いた素性のうちどの素性が有用になりそうかを次章において検討する.

第8章 *SVM*分離平面との距離に基づく素性分析

本章では性能向上がより見られた Case2 において，拡充した素性のうち，どの素性が段落の順序推定に有用であるかを検出するため，*SVM*分離平面との距離に基づく素性分析により，段落の順序推定に有用な素性を検討する．*SVM*分離平面との距離に基づく素性分析の手法を以下を示す．

8.1 手法

本節では，*SVM*分離平面との距離に基づく素性分析の手法について記載する．以下に素性分析の手順を示す．

1. 学習データで用いた個々の素性の 1 個のみの事例を持つテストデータを作成する
2. 1 で用いた学習データを学習データとして用いて 1 で作成したテストデータ内の全ての事例を *SVM*により分類する
3. 2 において，*SVM*で分類する際に各事例に対する分離平面との距離が算出される
4. 3 で得られた分離平面との距離により，分離平面との距離が大きい事例を段落の順序推定に有用な可能性の高い素性とする
5. 4 で得られた有用な可能性の高い素性を人手により検討する

1. 学習データで用いた個々の素性の 1 個のみの事例を持つテストデータを作成

実験に用いた学習データ内の各事例には多くの素性が付与されている．事例に付与された多くの素性を素性が 1 個のみの事例となるように全ての素性に対して作成し，これらをテストデータとする（図 8.1）．

学習データ内の各事例の形式は「どちらが先か@推定段落1 ☆推定段落2 素性」であり、データ内では1行につき1事例が付与されている。形式内の各項目について以下説明する。「どちらが先か」は事例に対する正解タグを示し、「推定段落1」及び「推定段落2」は入力となる推定する2段落の対を表し、段落1→2の順で入力に与えている。「@」、「☆」、「」（半角スペース）は各文字列の区切りを示し、以下素性は1個につき半角スペースでわけている。テストデータ内の形式は「どちらが先か@用いる素性 素性1個」であり、学習データの形式の各項目同様の部分は省略する。「@」以降の「用いる素性」はその事例に付与された「素性1個」を示す。

図8.1を例にすると、学習データ内の3事例に付与された10個の素性から各事例に付与された素性が1個のみとなる事例を10個作成する。

学習データ 事例形式：どちらが先か@推定段落1 ☆推定段落2 素性

段落A@段落A ☆段落B n1 n2 n3
段落A@段落B ☆段落A m1 m2 m3 m4
段落B@段落B ☆段落C o1 o2 o3



3事例から

段落A@n1 n1	段落B@o1 o1
段落A@n2 n2	段落B@o2 o2
段落A@n3 n3	段落B@o3 o3
段落A@m1 m1	
段落A@m2 m2	
段落A@m3 m3	
段落A@m4 m4	

10事例作成

テストデータ 事例形式：どちらが先か@用いる素性 素性1個

図 8.1: 学習データで用いた個々の素性を素性1個のみの事例を持つテストデータ作成

2. 1 で用いた学習データを用いて 1 で作成したテストデータ内の全ての事例を SVM により分類

1 で用いた元々の学習データを学習器に学習させ、1 で作成したテストデータ内の素性1個ずつの事例全てを SVM により分類する。

図8.2のように、図8.1で作成したテストデータとテストデータ作成の際に用いた元々の学習データを用いて機械学習によりテストデータ内の全事例（素性）を SVM に

よりどの段落に分類されるかを推定する。

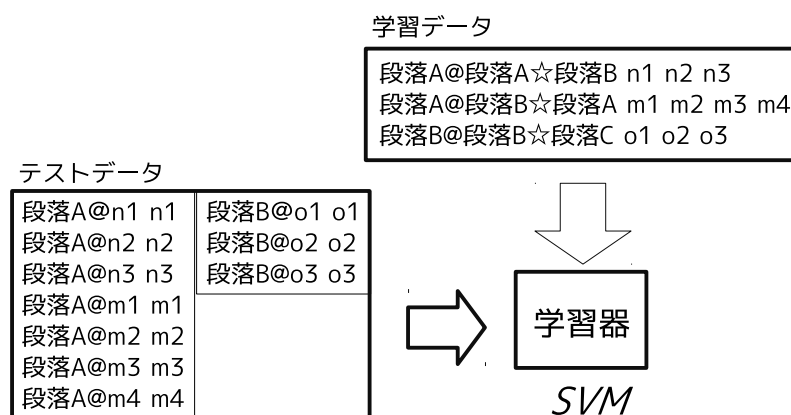


図 8.2: 1 の学習データを用いてテストデータを *SVM* により分類

3. 2 において、*SVM* で分類する際に各事例に対する分離平面との距離が算出される

SVM で分類する際に各事例（素性 1 個のみ）に対する分離平面との距離 d が算出される。

図 8.3 のように、*SVM* によりテストデータを推定する際に、学習データ内の全事例を学習することによりテストデータ内の各事例に対する分離平面との距離 d が算出される¹。

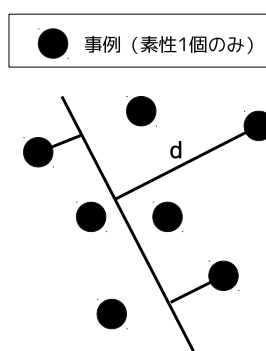


図 8.3: 各事例（素性 1 個のみ）に対する分離平面との距離

¹黒丸はテストデータ内の事例 1 個を示し、距離 d は事例と分離平面との最短距離を示す

4. 3で得られた分離平面との距離により，分離平面との距離が大きい事例を段落の順序推定に有用な可能性の高い素性とする

推定の際に分離平面との距離が各事例に対し算出され，テストデータ内の各事例と分離平面との距離が大きいほど段落の順序推定に有用な可能性の高い素性とする．

図 8.4 のように，推定結果の際に各事例と分離平面との距離が算出される．分離平面に離れるほどその事例に付与される素性は分類の際に有用な素性と考える．図では， d_3 の距離をもつ事例を有用な可能性の高い素性としている．

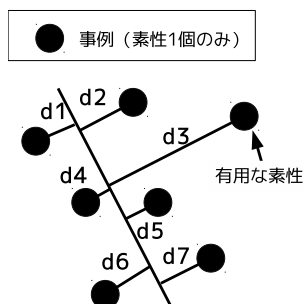


図 8.4: 分離平面との距離 d_n が大きい事例ほど有用な可能性の高い素性

8.1.1 4で得られた有用な可能性の高い素性を人手により検討する

各素性に対する分離平面との距離により得られた有用な可能性の高い素性に対し，人手によりデータ内の事例と見比べ，有用な素性を検討する．

8.2 実験条件

学習データは6章で用いた Case2 におけるデータ（毎日新聞 1992 年 7 月）を用いて分析を行う．本節で用いる学習データとテストデータの事例数を表 8.1 に示す．

表 8.1: 学習データとテストデータの事例数

	学習データ	テストデータ
データ数	15,020	111,877

有用な素性の検討方法として以下に示す．素性分析の結果は，1 行につき素性 1 個のみの事例とその事例との分離平面との距離とその素性の分類先が出力される．その出

力結果を分離平面との距離が大きい順になるようにソートし、ソートした結果内の分離平面との距離が大きい素性とデータ内の事例を人手により見比べ、段落の順序推定に有用な素性を検討する。

8.3 実験結果

素性分析手順3により得られた、各素性における分離平面との距離の結果を以下に示す。データ内の形式は「分類先, 分離平面との距離@付与された素性」である。“分類先”とは学習に基づきテストデータ内の各事例に“付与された素性”に対しての分類先を示す²。枠内の“R,0.329129@一文比率区間：0、5～1”を例に挙げると、一文比率区間（段落A→B、またはB→A間の隣接する各1文内に出現する単語の共起数をそれぞれ求め、求められた2つの値から比率を算出する）によって求められた値が0.5から1の間にある場合にR（段落B）に推定されやすいことを示している（数値である0.329129は用いた素性における分離平面との距離の値を示す）。

各素性における分離平面との距離の結果（一部）

R,0.329129@一文比率区間：0、5～1
L,0.319304@一文区間：0～1
L,0.317089@比率：0～
L,0.305045@P 一致数：1 0
L,0.296821@P 比率区間：0～0、5
L,0.296549@一文比率区間：0～0、5
L,0.295711@P 一致数：8
L,0.295643@LR 類似数 R ×: 1
L,0.295486@LR 類似数 R ×: 2
L,0.292465@P 一致数：3
L,0.290423@P 一致数：7
L,0.285407@P 類似数 L ○: 1
L,0.285366@L 日付け:1
R,0.285301@P 一文一致数：2
L,0.275282@R 日付け:0
L,0.275282@L 新名詞:日
R,0.273244@比率：1、5
L,0.262242@P 一致数：4
L,0.261228@P 一文一致数：3

また、人手により検討した。検討した結果、有用と思われた素性を表8.2に示す³。

²“,” や “@” を用いて分割するため、素性内の全角 “、” は表記ミスではない

³素性 ID は表5 参照

表 8.2: 人手による検討結果（推定段落 A, B）

素性	素性 ID	素性の説明	どの段落に分類する際に有用か
一文比率区間：0、5～1	b23	b23 で求めた値が 0.5 から 1 の間	段落 B が先になる場合
一文区間：0～1	b22	b22 で求めた値が 0 から 1 の間	段落 A が先になる場合
P 一致数：7	b28	b28 で求めた値が 7 以上	段落 A が先になる場合
L 日付け:1	a4	段落 A 内に日付を含む	段落 A が先になる場合
P 一文一致数：3	b24	b24 で求めた値が 3 以上	段落 A が先になる場合
P 比率：4	b29	b29 で求めた値が 4 以上	段落 A が先になる場合
LR 類似数 $R \times$ ：2～4	a7	a7 で求めた値が 2 から 4 の間	段落 B が先になる場合
P 類似数 $L \circ$ ：4～6	a14	a14 で求めた値が 4 から 6 の間	段落 A が先になる場合
L 特定品詞:1	a3	段落 A 内初めの文頭に連体詞または接続詞を含む	段落 B が先になる場合
R 名詞:同	a1	段落 B 内に名詞「同」を含む	段落 A が先になる場合

8.4 素性分析の考察

表 8.2 より，新聞記事を用いた段落の順序推定において有用な素性は，以前の段落と推定段落間との各 1 文の類似性や推定段落間の各 1 文の単語の共起数や以前の段落と推定段落間との段落内の単語の出現位置を用いた単語の共起数のように本研究で拡充した素性が有用な素性として検討した．また，先行研究で得られた素性分析同様，段落内の一方に日付けを表現する単語や特定の品詞を含む場合に段落の順序推定に有用な素性として検討した．

第9章 追加実験

本章では、追加実験として以下の2種類の実験を行う。

1. 段落間の隣接する近辺の細かな情報の類似性による推定法
2. 小説を用いた段落の順序推定

9.1 段落間の隣接する近辺の細かな情報の類似性による推定法

9.1.1 はじめに

本研究では、段落間の隣接する近辺の細かな情報（以下隣接情報と表記する）に着目し段落内の最初の文と最後の文の単語の共起数を素性として新たに組み込んでいる。本節ではこの素性に用いた段落間の隣接情報のみを用いて段落の順序を推定する。

9.1.2 推定方法

本節での推定方法を以下に示す。推定段落 A, B に対し、段落 $A \rightarrow B$ の順序で A の最後の文と B の最初の文との単語の共起数 X を求め、逆順の場合の単語の共起数 Y も同様に求める。求められた2つの値 X, Y のうち、より値が大きい順序間の内容が似通っていると判断し、順序に対して出力をする（図 9.1）¹。

9.1.3 実験条件

本節で用いる学習データとテストデータは表 7.1, 表 7.2 を用いる。実験では、素性拡充後と隣接情報のみによる推定法と比較手法の正解率を比較する。

¹ X, Y が同じ値の場合、 $A \rightarrow B$ の順序に対して出力をする

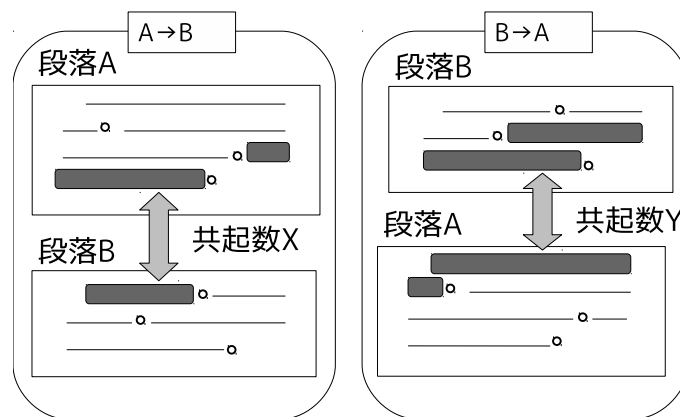


図 9.1: 推定方法

9.1.4 実験結果

表 9.1: 素性拡充後における提案手法と隣接情報のみによる推定法と比較手法の正解率

入力段落対	素性拡充後	隣接情報のみ	比較手法
Case1	0.853	0.595	-
Case2	0.619	0.585	0.528
Case3	0.676	0.607	0.618

表 9.1 より，段落間の隣接情報の類似性による推定法は全 Case において本研究で用いた手法より低い正解率であることがわかる．また，隣接情報の類似性による推定法は比較手法より高い正解率であることがわかる．

9.1.5 実験の考察

本研究の手法と比較すると，隣接する近辺の細かな情報のみによる推定法は全体的に低い性能であった．段落の初めに前段落の最後の内容が出現しない場合に対して，機械学習は段落間の隣接する近辺の細かな情報でなく他の特徴を用いて推定するため，性能に差があったと思われる．しかし，段落間の隣接する近辺の細かな情報の類似性による推定法で 0.5 以上の正解率を得ていることにより，この推定法は少なからず有用な手法であることがわかった．

9.2 小説を用いた段落の順序推定

9.2.1 はじめに

文章の種類によって、どのような違いがあるかを確認するために本節では小説を取り上げる。小説の順序推定に有用な表現が獲得できたなら、執筆の手助けができると思われる。

9.2.2 問題設定

本実験の問題設定を以下に示す。章内のある箇所まで段落の順序が確定しているとする。それより後の箇所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。確定している段落の次に来る箇所に対し、一方は正解、もう一方は不正解となる2段落を与え、その2段落のうちどちらが先になるかを推定する。順序を用いることのできる情報は、推定対象の2段落と確定している段落とする。

9.2.3 提案手法

現在、青空文庫を用いる研究は多く提案されているが、文章の順序推定に関する研究はなされていない。本実験では、小説に対して機械学習を行い、小説内の段落の順序推定を行うこととする。扱う対象は2段落を対としたものとし、その2段落対が入力された場合、どちらが先かを機械学習により推定する。

9.2.4 提案手法に用いる素性

本節では5章と同じ素性を用いる。

9.2.5 実験条件

本実験で用いる小説は、青空文庫[11]上のものを利用する。青空文庫とは、誰でもアクセスできる自由な電子本をインターネット上に集めようとする活動で、著作権の消失した作品と、自由に読んでもらっても構わないとされたものを電子テキスト化した上で揃えている。

学習データとテストデータに用いる作品の詳細を表9.2に示す。

表 9.2: 各データに用いる作品とその詳細

データ名	作品名	作成日	総段落数	総章数
学習データ	我輩は猫である	1905 年 1 月–1906 年 8 月	2,245 段落	11 章
テストデータ	坊ちゃん	1906 年 4 月	471 段落	11 章

また，実験で順序推定を行う際に用いる 2 段落対の組には 2 種類の場合を考慮して，学習データとテストデータを作成する．

1. 章内の最初の 2 段落のみの対の順序を推定する場合（以下 CaseI と表記）
2. 章内のあらゆる接続する 2 段落の対の順序を推定する場合（以下 CaseII と表記）

各場合での学習データとテストデータの組数を表 9.3 に示す．

表 9.3: 各場合における学習データとテストデータの組数

各場合	学習データ	テストデータ
CaseI	22	22
CaseII	4,468	920

9.2.6 実験結果

各場合における提案手法の正解率を表 9.4 に示す．

表 9.4: 各場合における提案手法の正解率

各場合	提案手法
CaseI	0.455
CaseII	0.535

9.2.7 実験の考察

表 9.4 を見ると，正解率が全体的に低く，7.3.1 節内の新聞記事を用いた場合に比べて大きな性能差が見られた．

小説内の段落の順序推定において、2 値分類にかかわらず CaseI で正解率がランダムの場合の正解率の 0.5 より低い値になるのは学習データの不足と考える。CaseI は章内の先頭 2 段落のみの対を用いてデータを作成するため、1 つの章からは正順と逆順の 2 組の事例が作成される。このため、11 章からなる作品を 1 つ用いた場合に 22 事例のデータしか扱うことができない。この学習データの不足に対する解決法として、複数の作品を用いる場合がある。5 つの作品を用いることで 100 事例以上の学習データが作成可能となり、用いる作品を増やすことにより学習データの不足を補うことができると思われる。

また、小説を用いた場合の提案手法の正解率が全体的に低い値であることは小説を用いた段落の順序推定では新聞記事での特徴が有用でなかったと考える。新聞は記事内で内容が完結するのに対し、小説は章ごとに流れがあるため、小説を用いる場合は前の章の情報も特徴として用いることで、章内の段落の順序を推定できると思われる。

第10章 今後の課題

10.1 機械学習に基づく段落の順序推定における素性の拡充

新聞記事は校正技術者により読み手に最も読みやすい文章にされているが、与えられた2段落のうちどちらを先に読んでも読み手が内容の理解できる場合がある。本実験では、新聞記事内のあった順序を正しい順序（正解）として評価しているため、段落の順序を推定する読み手は与えられた2段落のうちどの段落が先にくるかの2択問題を推定しており、上記の場合を考慮していないという問題がある。この問題により、連続2段落対の順序推定の場合の性能が機械学習、人手推定ともに低い結果となったと考えられる。問題解決のために与えられる2段落の順序がどちらでもいい場合に対して人手で正解を設定することを今後の課題として挙げる。

10.2 小説を用いた段落の順序推定

11章から構成される小説を用いたことにより、学習データの不足から正解率が0.5を下回るという問題が生じた。問題解決のため、複数の作品を用いることにより学習データの不足を補うことができると思われる。

また、本実験では新聞記事を用いた場合の特徴を用いたことにより、提案手法の正解率が低い結果となった。新聞とは異なり、小説は章ごとに流れがあるため前章の情報も特徴として用いることで、章内の段落の順序を推定できると思われる。そこで、人間が順序推定を行い、順序推定に有用となる特徴を検討することを今後の課題として挙げる。

10.3 上記以外

本研究では，新聞記事を用いて段落の順序推定を行い，人手による順序推定に近い性能を得た．段落の順序推定において，文章の種類によってどのような違いがあるかを確認するために小説を用いて実験を行い，学習データの不足や用いた特徴が有用でないことがわかった．日記や Wikipedia，論文のような他の分野で段落の順序推定をすることを今後の課題として挙げる．

第11章 おわりに

文章を元の順序通りに推定する研究のひとつである伊藤らの研究 [1] では、段落単位での類似性に着目し、順序推定において段落全体の類似性を素性として用いている。しかし、段落間の最後の文と最初の文の類似性である、段落間の隣接する近辺の細かな情報は用いていないという問題がある。段落間の隣接する近辺の細かな情報も用いることで段落の順序推定がしやすくなると考え、本研究では、機械学習に基づく段落の順序推定に、本研究では段落間の隣接する近辺の細かな情報として、段落間の最後の文と最初の文に着目し、伊藤らの素性に拡充する。

段落の順序を推定する実験において、記事内の接続する全ての2段落における順序推定では、段落間の隣接する近辺の細かな情報を素性として拡充することにより、素性を拡充する前 (0.60) よりと高い性能 (0.64) が得られ、人手による順序推定の性能 (0.66) に大きく近づいた。また、素性の拡充による性能向上がより見られた接続2段落対の順序推定の場合に対し、分離平面との距離に基づく素性分析を行った。人手で検討した結果、推定対象の段落間の各文内の単語の共起数を用いた素性や推定対象の各段落と以前の段落との段落間の単語の共起数を用いた素性が有用な素性として検出された。また、本研究での正解を新聞記事内の元々の順序としたため、どちらの順序でも読みやすい場合を考慮していない。これにより、人手による順序推定、機械学習ともに低い性能となってしまう問題がある。上記の問題解決のために、与えられる2段落の順序がどちらでも良い場合に対して人手で正解を設定することを今後の課題として挙げる。

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，細部に渡るご指導を頂きました，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，ご指導，助言を頂きました，村上仁一准教授，徳久雅人講師に心から御礼申し上げます．また，ご多忙の中，助言をいただきました木村周平教授に厚く御礼申し上げます．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] 伊藤聡史, 村田真樹, 徳久雅人, 馬青. 機械学習を用いた段落の順序推定実験. FIT2013 第 12 回情報科学技術フォーラム, pp. 215–216, 2013.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] TinySVM: <http://chasen.org/~taku/software/tinysvm/>.
- [9] 村田真樹, 井佐原均. 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. 電子情報通信学会技術研究報告 2001-NL-144, pp. 15–22, 2001.
- [10] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸. 形態素解析システム『茶筌』version2.2.5 使用説明書. Technical report, 奈良先端科学技術大学院大学, 2001.

[11] 青空文庫: <http://www.aorora.gr.jp/>.