概要

文章の生成や推敲の問題の一つに，文の順序推定がある．複数の文からなる文章の作成の際，わかりやすくなるようにそれらの文を適切な順序に並べる必要がある．

文の順序推定に関する研究の多くは文章要約の一環として行われており，要約前の文章から得られる情報を用いて文の順序推定を行うのが主な手法である．もし要約前の文章から得られる情報を用いずに文の順序推定が可能ならば，文生成における文の順序推定技術の応用範囲が広がる．そこで，本研究では，要約前の文章の情報を用いない文の順序推定の問題を扱う．

要約前の文章の情報を用いずに文の順序を推定する研究に関しては，確率モデルなどがあるが，教師あり機械学習により文の順序を推定する研究はなされていない．そこで文の順序推定の手始めとして，本研究では段落内の 2 文について，そのどちらを先に書くべきかを教師あり機械学習を用いて推定した．

その結果，提案手法の正解率 (0.72 から 0.77) は従来手法に基づく確率手法の正解率 (0.58 から 0.61) よりも高く，有用であることを確認した．

目 次

表 目 次

図 目 次

第1章　はじめに

文章の生成や推敲の問題の一つに，文の順序推定がある．複数の文からなる文章の作成の際，わかりやすくなるようにそれらの文を適切な順序に並べる必要がある．文の順序推定とは，複数の文について適切な順序を推定することである．

文の順序推定に関する研究の多くは文章要約の一環として行われており，要約前の文章から得られる情報を用いて文の順序推定を行うのが主な手法である [1]．もし要約前の文章から得られる情報を用いずに文の順序推定が可能ならば，文生成における文の順序推定技術の応用範囲が広がる．例えば，要約前の文章の情報を用いずに文の順序推定ができれば，その技術は文章の推敲にも利用できる．そこで，本研究では，要約前の文章の情報を用いない文の順序推定の問題を扱う．

要約前の文章の情報を用いずに文の順序を推定する研究に関しては，Lapata[2] の提案する確率モデルなどがあるが，教師あり機械学習により文の順序を推定する研究はなされていない．そこで教師あり機械学習でこの問題を扱うこととした．本研究では，教師あり機械学習としてサポートベクトルマシン (SVM) を利用する[1]．

本研究では教師あり機械学習とともに多くの情報を利用した文の順序推定の方法を提案する．確率モデルは多くの情報を用いることが困難である．それに対して，教師あり機械学習では多くの素性を設定することで容易に多くの情報を用いることができる．提案手法は多くの情報を利用するため，既存の確率モデルよりも高い性能を出すことが期待される．

文の順序推定の研究の手始めとして，本稿では，シンプルな問題を設定する．複数の段落をまたがった現象は複雑であると考え，段落内の情報のみを用いて，段落内の 2 文について，そのどちらを先に書くべきかを推定することを本稿で扱う問題とする[2]．

[1]本稿では，SVM のように素性を多数利用可能な手法のみを教師あり機械学習と呼ぶ．確率モデルは，確率を教師データからもとめるため，教師あり機械学習と見ることもできるが，教師あり機械学習と呼ばない．

[2]文章全体での文の順序推定は，2 文の順序推定の組み合わせで処理可能と考える．

以下に本研究の主張点をまとめる．

- 本研究には，文の順序推定に初めて教師あり機械学習を用いたという新規性がある．
- 本研究の文の順序推定の問題 (2 文のどちらを先に書くべきかを求める) において，教師あり機械学習を用いた提案手法の正解率 (0.72 から 0.77) は，確率モデルに基づく従来手法の正解率 (0.58 から 0.61) に比べて高いことを確認した．提案手法は性能が高いという有用性がある．
- 教師あり機械学習を用いる提案手法は多くの素性 (情報) を容易に利用できる．素性をさらに増やすことでさらなる性能向上が期待できる．
- 教師あり機械学習を用いる提案手法では，素性を分析することで，文の順序推定において重要な素性 (情報) を知ることができる．実験において素性を分析した結果，順序を判定する 2 文において 2 文目の助詞「は」までの自立語と 1 文目の自立語の助詞「は」より後の自立語に同じ語がどの程度あるかを調べる素性が，文の順序推定において重要であることがわかった．

第2章　関連研究

関連研究には以下のものがある．

Lapata[2] は確率モデルによる文の並べ替えを提案している．1 文目に 1 文目の単語がある場合に 2 文目に 2 文目の単語がくる確率を求め，その総積から 1 文目がある場合の 2 文目の生起確率を算出し，その確率が最大となる文の並びを構築するものである．

大田ら [3] は確率モデルに加えて統計情報を利用して．文の連接しやすさと文の連接しにくさを求めて文の順序を推定する手法を提案した．文の連接のしやすさは，連続する 2 文間における単語の連接確率から求める．文の連接しにくさは，1 文章における 2 単語の共起情報と連続する 2 文における 2 単語の共起情報の差から求める．

岡崎ら [1] は複数の記事から抜き出された文の並べ替えを行っている．要約前の文章から得られる情報を用いた手法であり，要約前の文章の情報を用いない本研究とは大きく異なる手法である．順序を推定する対象となっている文が，要約前の文書でどのような環境にあるかに着目した手法である．複数の記事から抜き出された文を話題毎にグループ分けし，グループ毎に記事が記載された順に文を並べる．その後，ある文 1 の前提知識と考えられる文が，その文 1 の前に来るように，文の並び順を改善する．文 1 の要約前文書での文 1 の先行文と類似する文を，文 1 の前提知識と考えられる文とする．

また，文生成に関連して教師あり機械学習により語順を推定する研究に内元ら [4] の研究がある．我々の研究が文の順序の推定に教師あり機械学習を利用するのに対して，内元らは語順の推定に教師あり機械学習を利用する．内元らは，教師あり機械学習として最大エントロピー法を用いる．内元らは，文節の係り受け情報をもとに，語順を推定する．正しい語順はコーパス内での語順である．このため，語順に関わる学習データはコーパスから自動で構築でき，人手で学習データを作成する必要がない．学習データを人手で作成する必要がないことは，我々の研究でも同様である．我々の研究では，正しい文の順序はコーパス内での文の順序であるので，文の順序に関わる学習データをコーパスから自動で構築できる．

第3章　問題設定

本研究での問題設定は以下の通りである．

文書が段落の単位で入力され，段落の途中まで文の順序が確定しており，途中からの順序が不明であるとし，不明箇所のうちの 2 文の順序を推定する．

推定の際に利用できる情報は，判定する 2 文と，その 2 文を含む段落内のその 2 文の一方が出現するまでの文章とする．

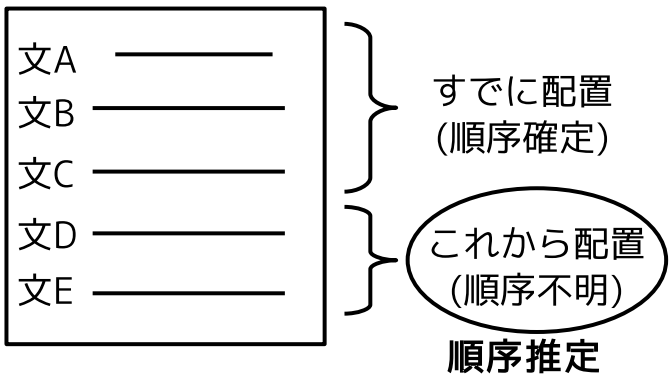

図 3.1: 問題設定の概略図

図 3.1 の場合，文 A から文 C までの順序推定が完了している (順序が確定している)．残り 2 文 (文 D,E) の順序が不明であり，本研究の問題設定では残り 2 文 (文 D,E) の順序を推定することになる．ここで使用できる情報は，推定対象である文 D,E の情報と，既に順序が確定している文 A かから文 C までの情報である．

第4章　提案手法

4.1　2文の順序推定方法

順序不明の2文 A, B が入力された場合，A-B の文の順序が正しいか否かを教師あり機械学習で判定する．正しい場合 A-B の順，そうでなければ B-A の順と推定する．本稿では，教師あり機械学習には SVM を利用する[1]．カーネル関数には2次の多項式カーネルを利用した．

4.2　データ作成

学習データは以下のようにして作成する．学習用の文章から2文を1組にして抜き出す．その2文から，元の文章通りの正しい順序 (正順) の2文とその逆の順序 (逆順) の2文を作成する．正順の2文を正例，逆順の2文を負例として，学習データを作成する．この時，正例に2文には正例であることを表すタグを付与する．負例の場合も同様に負例であることを表すタグを付与する．その後，文に含まれる情報を抜きだし，各素性を求める (素性については 4.3 節参照)．

テストデータも同様に，テスト用の文章から作成する．学習データ同様正例，もしくは負例のタグが付与されるが，テストデータの場合，これらのタグは SVM の出力が正しいかを評価するために用いる．

[1]具体的には，TinySVM[5] を用いる．

4.3 用いる素性

機械学習で用いられる個々の情報のことは素性と呼ばれる．教師あり機械学習法ではこの素性の設定が重要になる．本研究で用いた素性を表 4.1 に示す．ただし，素性が 2 文のうちの 1 文目と 2 文目のどちらで出現したかを区別する．1 文目の素性ならば「L」，2 文目の素性ならば「R」という接頭語を素性に付与して区別する．単語や品詞の取得には ChaSen[6] を用いる．各素性の詳細な説明を次ページより行う．

表 4.1: 素性

素性	説明
f1	文内で出現する単語とその品詞
f2	文内で出現する単語の品詞
f3	文の主語省略の有無
f4	文が体言止めで終わっているかいなか
f5	文内で最初に出現した助詞「は」で文を区切り，その前部で出現した単語とその品詞
f6	文内で最初に出現した助詞「は」で文を区切り，その後部で出現した単語とその品詞
f7	1 文目と 2 文目で使用されている助詞の対
f8	1 文目と 2 文目の単語の共起数
f9	1 文目においての f6 と 2 文目においての f5 が一致した度合い
f10	同じ段落内で，文の順序を判定する 2 文以前の文に出現する単語とその品詞
f11	同じ段落内で，文の順序を判定する 2 文の直前の文が体言止めで終わっているかいなか
f12	同じ段落内で，文の順序を判定する 2 文の直前の文の主語が省略されているかいなか
f13	同じ段落内で，文の順序を判定する 2 文の直前の文との自立語が一致した度合い

f1:文内で出現する単語とその品詞

f1 は文内で出現する単語とその品詞の組である．ただし，自立語か助詞「は」「が」「も」でないものは f1 としない．以下の素性の例では，コロンの前の表現は素性の種類を示す記号であり，コロンの後ろの表現はその素性が持つ情報である．1 文目の素性である場合は「L」という接頭語が付与され，2 文目の素性である場合は「R」という接頭語が付与される．

例

1 文目 この報告書は，国防総省の委託を受けた専門家グループがまとめた．

素性 L 名詞:報告 L 名詞:書 L 助詞:は L 固有名詞:国防総省 L 名詞:委託 L 動詞:受ける L 名詞:専門 L 名詞:家 L 名詞:グループ L 助詞:が L 動詞:まとめる

図 4.1: f1 の例

f2:文内で出現する単語の品詞

f2 は文内に出現する単語の品詞である．ただし，名詞の内で固有名詞であるものは固有名詞という情報をこの素性で用いる．助詞は f2 としない．

例

1 文目 この報告書は，国防総省の委託を受けた専門家グループがまとめた．

素性 L 品詞:名詞 L 品詞:固有名詞 L 品詞:動詞

図 4.2: f2 の例

f3:文の主語省略の有無

f3 は文の主語が省略されているかいなかである．文の主語が省略されていれば「1」，主語が省略されていなければ「0」とする．文中で助詞「は」「が」「も」が出現していなければ主語が省略されているとする．

例

1 文目 この報告書は，国防総省の委託を受けた専門家グループがまとめた．

素性 L 主語略:0

図 4.3: f3 の例

f4:文が体言止めで終わっているかいなか

f4 は文が体言止めで終わっているかいなかである．文が体言止めで終わっていれば「1」，体言止めで終わっていなければ「0」とする．文を後ろから検索し，記号以外で初めて出現した品詞が体言であれば体言止めであるとする．

例

1 文目 この報告書は，国防総省の委託を受けた専門家グループがまとめた．

素性 L 体止:0

図 4.4: f4 の例

f5:文内で最初に出現した助詞「は」で文を区切りその前部で出現した単語とその品詞

まず文中で最初に出現した助詞「は」で文を区切る．助詞「は」以前に出現する自立語の単語とその品詞を f5 とする．助詞「は」が存在しない場合，f5 は用いない．

例

1 文目 この報告書は，国防総省の委託を受けた専門家グループがまとめた．

素性 L 旧名詞:報告 L 旧名詞:書

図 4.5: f5 の例

f6:文内で最初に出現した助詞「は」で文を区切りその後部で出現した単語とその品詞

まず文中で最初に出現した助詞「は」で文を区切る．助詞「は」以後に出現する自立語の単語と品詞を f6 とする.「は」が存在しない場合，その文内の全ての自立語の単語と品詞を f6 とする．

例

1 文目 この報告書は，国防総省の委託を受けた専門家グループがまとめた．

素性 L 新固有名詞:国防総省 L 新名詞:委託 L 新動詞:受ける L 新名詞:専門 L 新名詞:家 L 新名詞:グループ L 新動詞:まとめる

図 4.6: f6 の例

f7:1 文目と 2 文目で使用されている助詞の対

f7 は 1 文目で使用されている助詞と 2 文目で使用されている助詞を対にしたものである．ただし,「は」「が」「も」以外の助詞は f7 としない．

コロンの後ろは，1 文目で使用されている助詞と 2 文目で使用されている助詞を連ねて表記している．例の場合，1 文目で「は」「が」が使用されていることを “L はが” 部分で表しており，2 文目でも「は」「が」が使用されていることを “R はが” 部分で表している．

例

1 文目 この報告書は，国防総省の委託を受けた専門家グループがまとめた．

2 文目 リーダーシップ拡散および内部紛争の下で，中国が分裂する可能性は五分五分と指摘している．

素性 対:L はが R はが

図 4.7: f7 の例

f8:1 文目と 2 文目の単語の共起数

f8 は 1 文目と 2 文目で共起した自立語の個数である．共起数を場合分けしたものも f8 として用いる．共起数の場合分けとして， 1 以上，2 以上，4 以上，6 以上，8 以上を用いる．

例

1 文目 私はサッカーが大好きだ．

2 文目 サッカーはとても面白いからだ．

素性 LR 類似度:1 LR 類似度:1~

図 4.8: f8 の例

f9:1 文目においての f5 と 2 文目においての f6 が一致した度合い

1 文目を Sa，2 文目を Sb とする場合を考える.「Sa→Sb」という順の 2 文で素性を考える場合，まず Sa を「は」で区切った後部にある自立語と，Sb を「は」で区切った前部にある自立語の一致数を求める．この一致数を A とする．Sa と Sb を逆にした場合でも同様に一致数を求め，それを B とする．$A - B$ の値を f9 とする．f8 と同様に，この値を場合分けしたものも f9 とする．$A - B$ が負となる場合も同様に場合分けし，それも f9 とする.

例

1 文目 私はサッカーが大好きだ.

2 文目 サッカーはとても面白いからだ.

素性 新旧類似度:1 新旧類似度:1~

図 4.9: f9 の例

f10:同じ段落内で文の順序を判定する 2 文以前の文に出現する単語とその品詞

f10 は段落の始めから文の順序を判定する 2 文までに存在する文に含まれる自立語とその品詞の組である．段落の始めの 2 文を判定する際は，それ以前の文が存在しないため f10 は用いない.

例

前文 私はサッカーが大好きだ.

1 文目 サッカーはとても面白いからだ.

2 文目 あのボールを蹴る感覚がたまらない.

素性 名詞:私 名詞:サッカー 名詞:大好き

図 4.10: f10 の例

f11:同じ段落内で文の順序を判定する 2 文の直前の文が体言止めで終わっているかいなか

f11 は，同じ段落内で，文の順序を判定する 2 文の直前の文での体言止めの有無の情報である．体言止めがされている場合「1」，されていない場合は「0」とする．段落の始めの 2 文を判定する際は，直前の文が存在しないため f11 は用いない．

例

前文 私はサッカーが大好きだ．

1 文目 サッカーはとても面白いからだ．

2 文目 あのボールを蹴る感覚がたまらない．

素性 体止:0

図 4.11: f11 の例

f12:同じ段落内で文の順序を判定する 2 文の直前の文の主語が省略されているかいなか

f12 は，同じ段落内で，文の順序を判定する 2 文の直前の文での主語省略の有無である．主語が省略されていれば「1」，主語の省略がされていなければ「0」とする．段落の始めの 2 文を判定する際は，直前の文が存在しないので f12 は用いない．

例

前文 私はサッカーが大好きだ．

1 文目 サッカーはとても面白いからだ．

2 文目 あのボールを蹴る感覚がたまらない．

素性 主語略:0

図 4.12: f12 の例

f13:同じ段落内で文の順序を判定する 2 文の直前の文との自立語が一致した度合い

文の順序を判定する 2 文として,「Sa→Sb」という順の 2 文 Sa,Sb が与えられ，この 2 文の直前の文を P とする．まず P と Sa の自立語の共起数を求め，それを α とする．同様に P と Sb の自立語の共起数を求め，それを β とする．α-β を f13 とする．また，f9 と同様の場合分けをしたものも f13 とする α-β が負となる場合も存在する．段落の最初の 2 文を判定する際は，直前の文が存在しないので f13 は用いない．

例

前文 私はサッカーが大好きだ.

1 文目 サッカーはとても面白いからだ.

2 文目 あのボールを蹴る感覚がたまらない.

素性　前文類似性:1 前文類似性:1~

図 4.13: f13 の例

第5章　確率手法

確率モデルに基づく従来手法と比較するために，確率手法でも文の順序推定を行う．確率手法とは，Lapata[2] の手法を参考にしたものであり，以下に確率手法の詳細を述べる．確率算出用の文書にある連接する 2 文から，それぞれの文に含まれる単語を抜き出す．1 文目の単語と 2 文目の単語のペアを作成し，1 文目に 1 文目の単語がある場合に 2 文目に 2 文目の単語がある生起確率を求める．そして，求めた生起確率の総積から 1 文目の文がある場合の 2 文目の文の生起確率 (以降，文の生起確率という) を算出する．本研究では 2 文の組において順序の推定を行うため，2 文から正順と逆順を作成し，正順の場合の文の生起確率と，逆順の場合の文の生起確率を求め，大きい方を正しい順番と推定する．$a_{\langle i,n\rangle}$ は文 S_i を構成する単語を表し，$a_{\langle i,j\rangle}$ と $a_{\langle i-1,k\rangle}$ が連接する 2 文に出現する確率は次式で表すことができる．

$$P(a_{\langle i,j\rangle}|a_{\langle i-1,k\rangle}) = \frac{f(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle})}{\sum_{a\langle i,j\rangle} f(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle})} \tag{5.1}$$

$f(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle})$ は単語 $a_{\langle i-1,k\rangle}$ がある文の次の文に単語 $a_{\langle i,j\rangle}$ が出現する頻度である．

第6章　実験

6.1　実験条件

機械学習に用いる学習データには，毎日新聞 91 年の 5 月分の記事を，評価に用いるテストデータには，毎日新聞 95 年 11 月の記事を用いる．確率手法での確率算出用の文書には，毎日新聞 91 年のすべての記事を用いる．

本研究では以下の 3 種類の実験を行う．

CASE1 段落内の最初の 2 文のみを用いる場合

CASE2 段落内全ての連接した 2 文を用いる場合

CASE3 段落内の全ての組み合わせで 2 文を用いる場合

各 CASE での学習データとテストデータの 2 文の組数を表 6.1 に示す．

表 6.1: 各 CASE での 2 文の組数

	CASE1	CASE2	CASE3
学習データ	33902	64290	130316
テストデータ	40386	82966	170376

6.1.1 CASE1:段落内の最初の 2 文のみを用いる場合

CASE1 は段落内の最初の 2 文のみを用いて，2 文 1 組を作成する．作成した組において，元の新聞記事内での順序を正しい順，その逆順を誤った順として学習データとテストデータを作成する．テストデータで作成した 2 つの順序のどちらが正しいかを教師あり機械学習を用いて推定する．推定結果が実際の新聞記事での順序であるかを確認する．CASE1 では，段落内において直前の文が存在しないため，素性 f10 から f13 は存在しない．

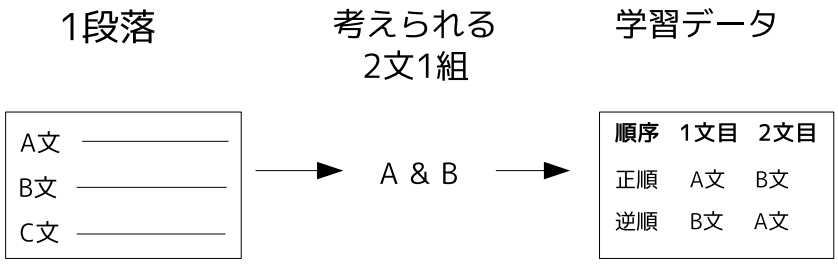

図 6.1: CASE1

図 6.1 の場合を例にあげる．文 A，B，C で構成される段落が与えられた場合，この段落の最初の 2 文，文 A と文 B の 2 文 1 組を作成する．元の記事での順序を (この場合文 A→ 文 B の順) を正例，その逆順 (この場合文 B→ 文 A の順) を負例として学習データを作成する．この時使用できる情報は，順序推定を行う 2 文から得られる情報である．テストデータも同様に作成するが，作成した文の順が正例か負例であるか，という情報は機械学習の出力が正しいかを評価する際に用いる．

6.1.2 CASE2:段落内の連接した2文を用いる場合

CASE2 は段落内の連接した 2 文を用いて，2 文 1 組を作成する．作成した組において，元の新聞記事内での順序を正しい順，その逆順を誤った順として学習データとテストデータを作成する．テストデータで作成した 2 つの順序のどちらが正しいかを教師あり機械学習を用いて推定する．推定結果が実際の新聞記事での順序であるかを確認する．

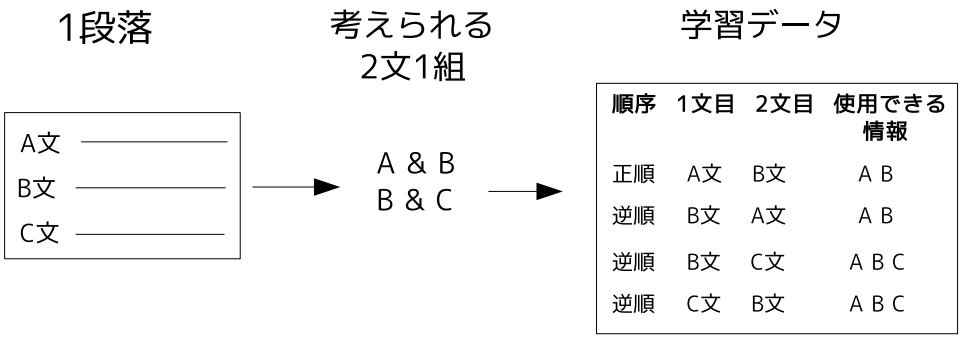

図 6.2: CASE2

図 6.2 の場合を例にあげる．文 A，B，C で構成される段落が与えられた場合，段落内の全ての連接した文の 2 文 1 組を考える．この場合考えられる組は，文 A と文 B，文 B と文 C の 2 通りである．それぞれの組において，元の記事での順序を (この場合文 A → 文 B の順，文 B→ 文 C の順) を正例，その逆順 (この場合文 B→ 文 A の順，文 C→ 文 B の順) を負例として学習データを作成する．使用できる情報は，順序推定を行う 2 文から得られる情報とそれ以前の文に出現した情報である．

6.1.3 CASE3:段落内の全ての文の組み合わせを用いる場合

CASE3 は段落内の文の全ての組み合わせを考慮し，2 文 1 組を作成する．作成した組において，元の新聞記事内での順序を正しい順，その逆順を誤った順として学習データとテストデータを作成する．テストデータで作成した 2 つの順序のどちらが正しいかを教師あり機械学習を用いて推定する．推定結果が実際の新聞記事での順序であるかを確認する．

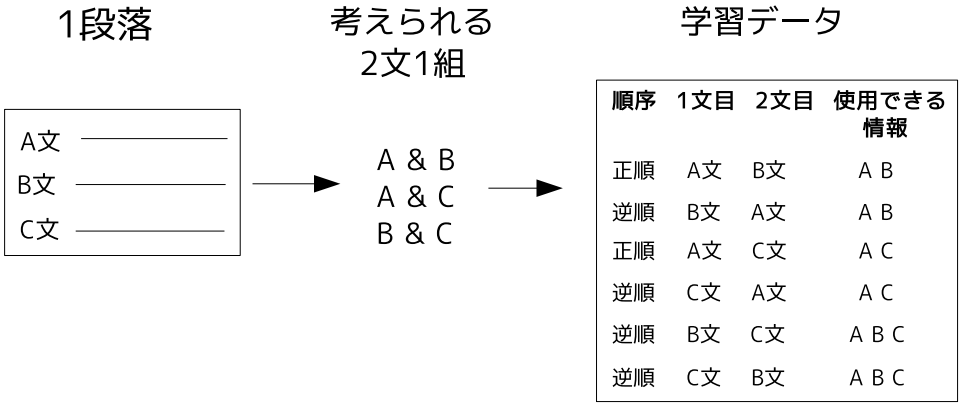

図 6.3: CASE3

図 6.3 の場合を例にあげる．文 A，B，C で構成される段落が与えられた場合，段落内の全ての文での組み合わせを考慮し，文の 2 文 1 組を考える．この場合考えられる組は，文 A と文 B，文 A と文 C，文 B と文 C の 3 通りである．それぞれの組において，元の記事での順序を (この場合，文 A→ 文 B の順，文 A→ 文 C の順，文 B→ 文 C の順) を正例，その逆の順序を (この場合，文 B→ 文 A の順，文 C→ 文 A の順，文 C→ 文 B の順) を負例として学習データを作成する．使用できる情報は，順序推定を行う 2 文から得られる情報とそれ以前の文に出現した情報である．

6.2 実験結果

6.2.1 提案手法と確率手法

提案手法と確率手法の正解率を表 6.2 に示す．確率手法では，2 文の順序を入れ替えたものと入れ替えないもので確率が同じになり，2 文の順序を推定できない場合がある．その場合は，その設問での正解の個数を 0.5 として正解率を計算している．表 6.2 のように，CASE1，CASE2，CASE3 ともに，提案手法の正解率 (0.72 から 0.77) が確率手法の正解率 (0.58 から 0.61) よりも高かった．

表 6.2: 正解率

機械学習			確率手法		
CASE1	CASE2	CASE3	CASE1	CASE2	CASE3
0.7677	0.7246	0.7250	0.6059	0.5835	0.5775

また，図 6.4 から図 6.9 にかけて，提案手法で正解したが確率手法で失敗した実例，その逆の実例を示す．

CASE1:提案手法○，確率手法×の実例

正解 文 1→ 文 2

文 1 小選挙区候補では，岩国哲人・前出雲市長が東京六区から出馬．

文 2 比例代表で公認決定していた現職の高橋一郎氏が東京七区に回った．

図 6.4: CASE1 での提案手法○，確率手法×の実例

CASE1:提案手法×，確率手法○の実例

正解 文 1→ 文 2

文 1 見直しを求められた点について，大和銀は「内容は言えない」（川上敏朗副頭取）としている．

文 2 これに対し，大蔵省は「改善点をもっと具体的に書くよう指示した」と話している．

図 6.5: CASE1 での提案手法×，確率手法○の実例

CASE2:提案手法○，確率手法×の実例

正解 文 1→ 文 2

前文 政府は三十一日，沖縄県の米軍基地の整理・統合問題を検討する新たな日米協議機関は，設置時に結論を出す期限を設けるよう米側に求める方針を固めた．
期限は「一年」とする案が有力になっている．

文 1 政府筋が明らかにした．

文 2 これは新たな基地の整理・統合であっても早期に決着させなければ，沖縄県民の不満が解消されない，との判断からで，米側の理解を求める方針だ．

図 6.6: CASE2 での提案手法○，確率手法×の実例

CASE2:提案手法×，確率○の実例

正解 文 1→ 文 2

前文 政府は三十一日，沖縄県の米軍基地の整理･統合問題を検討する新たな日米協議機関は，設置時に結論を出す期限を設けるよう米側に求める方針を固めた．

文 1 期限は「一年」とする案が有力になっている．

文 2 政府筋が明らかにした．

図 6.7: CASE2 での提案手法×，確率手法○の実例

CASE3:提案手法○，確率手法×の実例

正解 文 1→ 文 2

前文 総合的視野を小中学校時代から培うディベートを楽しむ学校教育が始まっていると知り，頼もしい限りだ．

文 1 論争するのは自分の論理に固執することではない．

文 2 相手の論理と比較対照し，より合理的な視野を確立していくために必要なことだ．

図 6.8: CASE3 での提案手法○，確率手法×の実例

CASE3:提案手法×，確率手法○の実例

正解 文 1→ 文 2

前文 自分の論理が絶対に正しいとは限らない．
相手の話を傾聴し，自分の論理に疑問が生じたり，間違いに気づくこともある．

文 1 また，時の流れとともに人間社会も刻々と変化し，確立したはずの論理も現状と順応しなくなる．

文 2 永久不変ということはあり得ない．

図 6.9: CASE3 での提案手法×，確率手法○の実例

今回用いた確率手法では，学習新聞記事内に存在しない単語の組み合わせが出現した場合，学習新聞記事内で一番小さな確率の 1/100 倍の確率を与えている．しかし，先行研究においては，バックオフスムージングを行っており，厳密には本研究で比較用に用いたものとは少し異なっている．また，本研究での比較用の確率手法で用いる単語の生起確率は，段落内の全ての連接した 2 文から単語生起確率を求めている．しかし，行った実験は 3 種類であり，CASE 毎にテストデータを作成している．よって今後，各 CASE 毎に学習も作りかえてもう一度比較実験を行う必要があると考えている．

6.2.2 人手による文の順序推定の正解率との比較

毎日新聞 95 年 11 月分の記事から，CASE 毎にランダムに 100 組 (各組は 2 文からなる) を抜き出す．CASE1，CASE2，CASE3 を 1 人それぞれ 20 問ずつ，被験者 5 名で，その 100 組について，文の順序を推定する．同じ 100 組を提案手法と確率手法で文の順序を推定する．CASE2，CASE3 において機械学習では判定する 2 文以前の (同じ段落内の) 文の情報を用いているので，人の判定の際でも判定する 2 文以前の (同じ段落内の) 文を提示する．

表 6.3 に被験者と提案手法と確率手法による判定結果の正解率を示す．表の A から E は被験者 5 名を，平均は被験者 5 名の正解率の平均を意味する．

表 6.3: 人による文の順序推定の正解率との比較

	被験者						提案	確率
	A	B	C	D	E	平均	手法	手法
CASE1	0.75	0.70	0.75	0.95	0.95	0.82	0.79	0.65
CASE2	0.80	0.80	0.85	1.00	0.90	0.87	0.67	0.64
CASE3	0.65	0.75	0.85	0.65	0.70	0.72	0.71	0.56

表 6.3 の被験者の正解率の平均と提案手法の正解率を比較すると，CASE1 と CASE3 では，提案手法は被験者の平均に近い正解率を得ている．CASE2 では，残念ながら提案手法は被験者の平均よりかなり低い正解率となっている．CASE2 に関して提案手法で設定した素性がまだ不十分かもしれない．今後素性を拡充し，被験者の正解率に近づけたいと考えている．

図 6.10 から図 6.15 にかけて各 CASE 毎の，提案手法で推定成功し人手で推定失敗した実例，またその逆の実例を示す．

CASE1:提案手法○，人手推定×の実例

正解 文 1→ 文 2

文 1 しかし，首相自身の日刊併合に関する発言で両国間に緊張状態が起きている中で江藤長官の発言問題が表面化し，韓国が過敏になっていることを首相官邸は甘くみたといえそうだ.

文 2 日韓問題が村山政権を揺さぶる可能性が強まってきた.

図 6.10: CASE1 での提案手法○，人手推定×の実例

CASE1:提案手法×，人手推定○の実例

正解 文 1→ 文 2

文 1 それだけではない.

文 2 秋には予備校が実施する東大用模試の会場に出向き，めぼしをつけた受験生にアタック.

図 6.11: CASE1 での提案手法×，人手推定○の実例

CASE2:提案手法○，人手推定×の実例

正解 文 2→ 文 1

前文 一つはイスラム復興運動だ．イスラム圏では一九七○年代後半,「宗教の世俗化」（宗教と政治の分離）に反対するイランのホメイニ革命が弾みとなって,「政教一致」を前面に打ち出す「原理主義」が復活した．

文 1 このため「原理主義」派からターゲットにされている．

文 2 トルコはエジプトと並び中東イスラム圏では珍しい世俗国家だ．

図 6.12: CASE2 での提案手法○，人手推定×の実例

CASE2:提案手法×，人手推定○の実例

正解 文 2→ 文 1

前文 農林系金融機関が「被害者」と主張することも的はずれだ．農民からカネを預かって運用を任されている以上,「だまされた」では済まない．

文 1 「元本ロスは認められない」との主張が最後まで通るのかは疑問だ．

文 2 農林系は，フタを開ければ，住専だけではなく，別の不動産融資への焦げ付きが表に出ることを恐れているといわれる．

図 6.13: CASE2 での提案手法×，人手推定○の実例

CASE3:提案手法○，人手推定×の実例

正解 文 2→ 文 1

前文 藤ノ木古墳法隆寺の西約三百五十メートルにある六世紀後半の円墳.

文 1 直径四十八メートル，高さ九メートルの墳丘部を含む一帯は国の史跡，石棺内外の遺物も国の重要文化財に指定されている.

文 2 一九八五年の第一次調査で，石室内から最高級の馬具が出土.

図 6.14: CASE3 での提案手法○，人手推定×の実例

CASE3:提案手法×，人手推定○の実例

正解 文 2→ 文 1

前文 岐阜県大野郡清見村の彦左衛門，といっても人間ではない.

文 1 一つの大きな成果が出るまでには長い時間がかかる.

文 2 飛騨の山奥に立つ樹齢約九百八十年のミズナラの大木である.

図 6.15: CASE3 での提案手法×，人手推定○の実例

6.3 素性の分析

本研究で使用した素性のうちどの素性が文の順序推定に有用かを確認する．具体的には，CASE3 において，素性を一つ取り除いた場合と，素性を全て使用した場合の正解率を比較する．素性を一つ取り除いた場合の正解率と，全ての素性を用いた場合の正解率との差を表 6.4 に示す．

表 6.4: 素性を取り除いた場合の正解率

取り除いた素性	正解率	差
f1	0.7211	-0.0039
f2	0.7226	-0.0024
f3	0.7251	+0.0001
f4	0.7251	+0.0001
f5	0.7212	-0.0038
f6	0.7223	-0.0027
f7	0.7243	-0.0007
f8	0.7201	-0.0049
f9	0.6587	-0.0663
f10	0.7172	-0.0078
f11	0.7240	-0.0010
f12	0.7241	-0.0009
f13	0.7241	-0.0009

表 6.4 のように，f9 の素性を使用しない場合正解率が極端に落ちることがわかる．f9 が特に文の順序推定において重要であることがわかる．f9 を使用すると推定に成功するが，f9 を使用しないと推定に失敗する例を以下に示す．

f9 が成否に関わる例

文 1:そこで，ドルを使ったのが始まりだ．
文 2:その後一ドルは三六〇円から一〇〇円まで下がり続けた．

正解の文順は「文 1→ 文 2」であるのに対し，f9 を使用しない場合「文 2→ 文 1」であると推定した．f9 の素性は 2 文目の助詞「は」までの自立語と 1 文目の自立語 (1 文目に助詞「は」がない場合) に同じ語があるかを調べるものである．上記の例文では単語「ドル」が 2 文目の助詞「は」までと 1 文目の両方に存在するため，f9 により正しく文の順序を推定できたものと思われる．

第7章　今後の課題

本稿では，段落内の情報しか扱わなかった．しかし，文章全体での文の並べ替えを対象とした場合は，段落の外の情報も利用していくべきと考える．また，複数の段落をまたがった2文での文の順序推定や段落の順序の推定も考慮する必要がある．今後はこれらの事柄も扱っていきたい．

また，素性の拡充，もしくは改良を行うことでさらに性能向上を目指したいと考えている．

第8章　おわりに

本研究では，文の順序の推定に教師あり機械学習を用いる新規な手法を提案した．文の順序を推定する実験において，提案手法の正解率 (0.72 から 0.77) は従来手法に基づく確率手法の正解率 (0.58 から 0.61) よりも高かった．また，人手での順序推定との比較の結果，CASE1,CASE3 においては人手評価に近い正解率 (0.01 から 0.03 程度の差異) を得られた．これらのことから，本研究での提案手法は有用であることごが確認できた．素性を分析したところ，2 文目の助詞「は」までの自立語と 1 文目の自立語の助詞「は」より後の自立語に同じ語がどの程度あるかを調べる素性を取り除き順序推定を行った場合の正解率が一番減少した (0.72 から 0.65 に減少) ことを確認した．このことから本研究で用いた素性の中では，2 文の新情報，旧情報に着目した素性が文の順序推定に最も有効であることがわかった．

謝辞

本研究を進めるに当たり，終始に渡り研究の進め方や本論文の書き方など，細部にわた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講座 C 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，村上仁一准教授，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いた計算機工学講座 C 研究室の皆様に感謝の意を表します．

参考文献

[1] 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.

[2] Mirella Lapata: “Probablistic Text Structuring: Experiments with Sentence Ordering”，In Proceedings of the 41st Meeting of the Association of Computational Linguistics，pp.545-552，2003.

[3] 大田 浩志，山本 和英: “文書生成のための文の並べ替え”，言語処理学会 第 15 回年次大会 発表論文集，pp.813-816，2009.

[4] 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”，言語処理学会誌 (自然言語処理)，Vol.7，No.4，pp.163-180，2000.

[5] TinySvm : http://chasen.org/taku/software/TinySVM/

[6] ChaSen : http://chasen-leagacy.sourceforge.jp/