教師あり機械学習を用いた文の順序推定

林裕哉 †　　村田真樹 ‡　　徳久雅人 ‡
† 鳥取大学 工学部 知能情報工学科
‡ 鳥取大学 大学院 工学研究科 情報エレクトロニクス専攻
{s082043, murata, tokuhisa}@ike.tottori-u.ac.jp

1　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．

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．

2　関連研究

文の順序に関する関連研究には Lapata[2] の研究の他に，大田ら [3] の研究がある．大田らの研究は，Lapata の手法に類似する．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．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，

*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．

表 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

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．

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は以下の通りである．順序を推定する 2 文が順序付きで与えられる．与えられた 2 文の順序が正しいかどうかを推定する．

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．

3.2 提案手法

教師あり機械学習を利用して，順序付きで与えられた 2 文の順序が正しいかどうかを推定する．本稿では，教師あり機械学習には SVM を利用する[*3]．カーネル関数には 2 次の多項式カーネルを利用した．

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．

順序の推定は以下のようにして行う．順序を推定する 2 文が順序付きで与えられた場合，教師あり機械学習でそれが正例であるか負例であるかを推定する．正例と推定された場合，その 2 文の順序は正しいと判定し，負例と推定された場合，その 2 文の順序は正しくないと判定する．

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 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」という印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以下で行う．

[*3] 具体的には，TinySVM[5] を用いる．

3.3.1 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．

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

3.3.2 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 名詞 L 固有名詞 L 動詞

3.3.3 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

3.3.4 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

3.3.5 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

3.3.6 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

3.3.7 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である．ただし，「は」「が」「も」以外の助詞は f7 としない．

例:(本文 1) 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．
(本文 2) 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

3.3.8 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:(本文 1) 私はサッカーが大好きだ.
(本文 2) 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~

3.3.9 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:(本文 1) 私はサッカーが大好きだ.
(本文 2) 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~

3.3.10 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の単語と品詞である．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.
(判定文 1) サッカーはとても面白いからだ.
(判定文 2) 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

3.3.11 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.
(判定文 1) サッカーはとても面白いからだ.
(判定文 2) あのボールを蹴る感覚がたまらない.
(素性) 体止:0

3.3.12 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.
(判定文 1) サッカーはとても面白いからだ.
(判定文 2) あのボールを蹴る感覚がたまらない.
(素性) 主語略:0

3.3.13 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．α-β が負となる場合も存在する．f9 と同様の場合分けをする．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.
(判定文 1) サッカーはとても面白いからだ.
(判定文 2) 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~

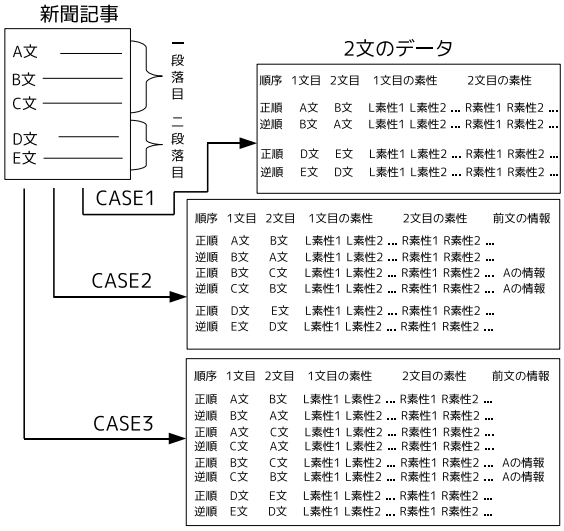

図 1: 2 文 (1 組) の取り出し方

4 確率手法 (比較手法)

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．

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

$f(a_{\langle i,j\rangle}, a_{\langle i-1,k\rangle})$ 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．

5 実験

5.1 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．

実験で用いる 2 文には，以下の 3 種類のものを用いる．段落内の最初の 2 文のみを用いて 2 文の組を作成する場合 (CASE1)，段落内全ての連接した 2 文を用いて 2 文の組を作成する場合 (CASE2)，段落内の全ての 2 文の組み合わせを考慮し，それらすべてから 2 文の組を作成する場合 (CASE3) である (図 1 参照)．上記とするため，CASE1 では，判定する 2 文以前の文が存在せず，f10 から f13 までの素性が存在しない．

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

表 2: 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

表 3: 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

表 4: 人による文の順序推定の正解率との比較

	被験者						提案手法	確率手法
	A	B	C	D	E	平均		
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

5.2 実験結果

提案手法と確率手法の正解率を表 3 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 3 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．

5.3 人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 4 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 4 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．

5.4 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 5 に示す．

表 5 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．

表 5: 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

f9 を除いて失敗した例

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．

6 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．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が最も有効であることがわかった．

今後は素性を拡充することでさらに性能向上を目指したいと考えている．

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．

謝辞

本研究は科研費（23500178）の助成を受けたものである。

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/