E-015

機械学習を用いた段落の順序推定実験

Experiments on Order Estimation of Paragraphs by Machine Learning

伊藤 聡史[*1] 村田 真樹[*1] 徳久 雅人[*1] 馬 青[*2]
Satoshi Ito Masaki Murata Masato Tokuhisa Qing Ma

1 はじめに

人が文章作成を行う際，読者が読みづらい文章を作成することがある．読みづらい文章には，意味の分からない専門用語を用いることや，狭い文章中に複数の話題が存在すること，冗長な文章を用いること，文章の順番が良くないことなど，様々な原因が存在する．本研究では，そのうちの文章の順番の問題を取り上げる．文章の順番が良くないために読みづらい文になっている場合は，文章を適切な順序に並べ替える必要がある．文章を適切な順序に並べ替えるために，本研究では機械学習を利用する．機械学習には性能が高いと広く認識されている *Support Vector Machine*(*SVM*) を用いる．

本研究では，2 段落ごとで元の順番 (正順)・その逆の順番 (逆順) の 2 通りについての問題を作成し，機械学習を用いることにより，どちらの順序が正しいかを判定する．

2 関連研究

林らは新聞記事から文の順序推定のために，多数の素性を用いた教師あり機械学習に基づく研究を行った [1]．新聞記事から 2 文一組で抜き出し，その 2 文から元の順の文（正例）と逆順の文（負例）を作成し教師あり機械学習を用いてその 2 文が正例か負例かを判定して文の順序を推定するというものである．機械学習に用いるデータは，内元らの研究 [2] を参考にしてコーパスから自動で構築できるようにした．実験では，段落内最初の 2 文のみを用いる場合と，段落内全ての連接した 2 文を用いる場合と，段落内全てから 2 文を用いる場合の 3 種類における順序推定を行った．

伊藤ら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教師あり機械学習に基づく研究を行った [3]．林ら同様新聞記事から 2 段落 1 組で抜き出し，その 2 段落対から元の順（正順）と逆順の問題を作成し教師あり機械学習を用いてその 2 段落対が正しい順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と，記事内全ての連接した 2 段落を用いる場合の 2 種類における順序推定を行った．

これらの研究は文章の順序推定を行っている．本稿はを伊藤らの研究 [3] とほとんどよく似た継続研究である．伊藤らと本稿との違いは記事内あらゆる 2 段落対の組み合わせの場合の実験を増やしたことにある．

Danushka らは複数文書からの要約作成のために文の順序推定の研究を行った [4]．文の順序推定には，時間的情報，内容の意味的近さ，要約前文章での文の順序などの情報を素性とした教師あり機械学習法を用いた．

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

3 問題設定と提案手法

3.1 問題設定

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

3.2 提案手法

段落の順序を推定する 2 段落が順序付き (正順または逆順) で入力された場合，その順序が正解の順序と同じ順序か否かを機械学習により判定する．機械学習としては *SVM* を利用する．*SVM* には，*TinySVM* を利用する [5]．カーネル関数には 2 次の多項式カーネルを利用する．

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

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

3.3 提案手法で用いる素性

機械学習で用いられる識別用の情報のことを素性といい，機械学習は与えられたデータを用いて上手く識別できるような素性を学習する．単語や品詞の情報の取得には，形態素解析システムの *ChaSen*[6] を用いる．本研究に用いる素性を以下に示す．素性の詳細は伊藤らの研究を参照されたい [3]．

素性 1,2 各段落に出現する単語の品詞情報
素性 3,4 『この＋名詞』や『同＋名詞』など特定の表現や『しかし』や『そして』など接続詞が出現するか否か
素性 5-11 推定を行う段落対の類似性
素性 12-19 以前の段落群と各段落の類似性
素性 20,21 新規単語が出現するか否か

[*1] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
[*2] 龍谷大学 理工学部 数理情報学科

4　比較手法

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

5　実験

5.1　実験条件

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

実験で用いる 2 段落の組は，以下の 3 種類の場合を考慮して作成する．記事内の最初の 2 段落のみを用いて作成する場合 (Case1), Case1 も含む記事内全ての連接する 2 段落を用いて作成する場合 (Case2), Case1,2 も含む記事内あらゆる 2 段落対を用いて作成する場合 (Case3) とする．但し, Case1 は先頭 2 段落のみを用いるため，比較手法の 4 節で挙げた比較手法は用いることができない．

Case1 は先頭の段落対であり推定する 2 段落以前の段落が存在しないので, 推定する 2 段落以前の段落情報を用いる素性 (素性 12 から 21) は用いない．

Case1 に用いる学習データの 2 段落対の組数は 1,550 組，Case2 の組数は 29,434 組, Case3 の組数は 80,248 組である．

5.2　提案手法と比較手法の比較実験

テストデータは毎日新聞記事 1992 年 8 月 1 日の 1 日分から作成する．Case1 での段落対の組数は 418 組であり，Case2 での段落対の組数は 3,146 組, Case3 での段落対の組数は 7,374 組である．表 1 に提案手法と比較手法の正解率を示す．

表 1 の提案手法と比較手法を比較すると，Case1 での提案手法は比較手法はないが約 8 割という高い正解率をであり，Case2,3 での提案手法は比較手法より正解率が高いことが分かる．また Case3 は Case2 より全体的に高い正解率であることが分かる．

表 1: 提案手法と比較手法との正解率

	提案手法	比較手法
Case1	0.8517	
Case2	0.5976	0.5277
Case3	0.6511	0.6181

5.3　人手との比較実験

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

Case1 は毎日新聞記事 1993 年 6 月，Case2 は同年 7 月, Case3 は同年 8 月, それぞれ 1 ヶ月分の記事からランダムに 2 段落対 1 組を 50 組を抜き出し，これらをテストデータとして用いる．

表 2 に提案手法と比較手法と被験者の正解率を示す．平均は，被験者の正解率の平均を示す．

表 2 の提案手法と比較手法と被験者を比較すると，Case1 では提案手法が被験者の平均の性能と同等であることから，Case1 での提案手法は人間と同程度の性能をもつことがわかる．また，Case2,3 をみると，提案手法は比較手法より高いが，被験者の平均よりも低い．全体的にみると，Case2 では提案手法だけでなく人手も 6 割と低い正解率であるため，Case2 のような連接している段落の順序推定は難しいと思われる．Case3 では比較手法を除き 7 割と Case2 の正解率を全て上回っている．

Case2 より Case3 の方が全体的に正解率が良い．この理由として以下が考えられる．Case3 は記事内あらゆる 2 段落対を用いるため，推定を行う段落対が非常に離れた個所にある 2 段落の場合がある．この場合は，以前とのつながりを利用して判断することでどちらが先に書くべき段落かを容易に判断できる場合があり，これにより Case3 の方が正解率が良かったと思われる．

表 2: 提案手法・比較手法・人手の順序推定の正解率

	提案手法	比較手法	被験者		
			A	B	平均
Case1	0.88		0.92	0.84	0.88
Case2	0.60	0.56	0.68	0.64	0.66
Case3	0.72	0.56	0.84	0.70	0.77

6　おわりに

本研究では段落の順序推定に機械学習を用いる手法を提案した．段落の順序を推定する実験において，記事先頭 2 段落の順序推定を行った場合提案手法は 0.85 という高い正解率を得た．人手による順序推定と同等の正解率であった．また，連接した 2 段落での順序推定では，提案手法は約 6 割という正解率であった．前方の文章との名詞の一致数が大きい方を前方とするベースライン手法よりは高い正解率であった．また，Case3 では比較手法を除き 7 割と Case2 の正解率を全て上回っていた．Case2 より Case3 の方が全体的に高い正解率であった．Case3 は推定を行う段落対が非常に離れた個所にある 2 段落の場合があるため，以前とのつながりを利用して判断することでどちらが先に書くべき段落かを容易に判断できたと思われる．

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] 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.
[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，自然言語処理研究会報告，2000(11)，pp. 55-62，2000.
[3] 伊藤 聡史，村田 真樹，徳久 雅人，馬 青: “教師あり機械学習を用いた段落の順序推定”，言語処理学会 第 19 回年次大会 発表論文集，pp. 442-445，2013.
[4] Danushka Bollegala, Naoaki Okazaki, Mitsuru Ishizuka: “A bottom-up approach to sentence ordering for multi-document summarization”, Information Processing & Management, Vol. 46, No.1, pp. 89-109, 2010.
[5] TinySvm : http://chasen.org/taku/software/TinySVM/
[6] ChaSen : http://chasen-leagacy.sourceforge.jp/