小説自動要約のための隣接文間の結束性判定手法

山本悠二 [1]　増山繁 [2]　酒井浩之 [3]

豊橋技術科学大学 知識情報工学系

[1,3] {yamamoto, sakai}@smlab.tutkie.tut.ac.jp
[2] masuyama@tutkie.tut.ac.jp

1　はじめに

近年，ウェブから膨大な量の機械可読な文章を入手することが可能となった．その中には，小説などの文学作品が多く含まれている．例えば，インターネット電子図書館として知られている青空文庫[*1]では，約 5,100[*2]もの作品を無料で入手することができる．しかし，これらの作品には要約が存在しないため，ユーザにとって新たにどの作品を読むかを決めることが困難である．そのため，作品を読むかどうかを決めるための要約 (指示的要約と呼ばれる) を自動的に生成する手法を確立することは重要であると考えられる．

小説を対象とした自動抄録の実験としては文献 [1] に収録されているものが挙げられる．これは話題展開の測定の研究では文章の中に新しい単語が出現することが話題展開の軸として重視されていることと，その他の文章における語の出現頻度に関する知見に基づいて，主に話題の提出・転換と話の結末と考えられる文を抽出するものである．

この文抽出による要約は話題の展開を考慮しているという点で有用な手法であると考えられる．しかし，指示詞やゼロ代名詞の照応が取れないといった難点がある．例えば，以下の文[*3]において二文目のみが抽出されるとする．

1. 小田原熱海間に、軽便鉄道敷設の工事が始まったのは、良平の八つの年だった。
2. 良平は毎日村外れへ、その 工事を見物に行った。

この場合，指示詞「その」が何を指しているのかを判断することができない．また，以下の文[*4]において二文目のみが抽出された場合，誰が動作を行っているのか判断することができない．

1. 余は子規の描いた画をたった一枚持っている。
2. 亡友の記念(かたみ)だと思って長い間それを袋の中に入れてしまっておいた。

いずれの場合も，要約として脈絡のないものとなり，内容をつかむことが難しくなるといえる．また，語の出現頻度による文抽出は，話題の展開のような文章の大域的な要素を考慮に入れることができるが，文間のつながりのような局所的な要素を考慮に入れることは難しい．そのため，より脈絡を考慮した要約を行うために文章内の整合性に大きく影響する言語的つながり (結束性) を考慮に入れることが考えられる．論説文に対して結束性を利用した要約手法としては文献 [2] などがある．一般に結束性を求める際には各文が前後の文とどのくらい結束しているかを定量化することが多い．しかし，結束していない文に対しても，文間の類似度が高いことや指示対象の推定の失敗により，結束度が高くなり，不適切な文抽出となる可能性がある．

そこで本稿ではある文とその前の文との結束性の有無に着目して要約を行う手法の提案を行う．

2　提案手法

2.1　考え方

隣接文間の結束性を用いてどのように要約するかを考える．まず，文体が一貫している文章で，ある隣接文間に結束性がない場合，その文間が何らかの意味の切れ目であると捉えることができる．そこで，隣接文の結束性の有無を判定し，結束性がないと判断したものを重要文として認定して文抽出を行う．また，行がえ，一字字下げによって明示的に文章のまとまりを示す形式段落に関しても作者が何らかの意図で意味の切れ目であると表していると考えることができるため，先頭の文を抽出する．

[*1] http://www.aozora.gr.jp/
[*2] 2006 年 2 月現在
[*3] 芥川龍之介『トロッコ』
[*4] 夏目漱石『子規の画』

Step 1. 文を順番に読み込む.
Step 2. 読み込んだ文が,
- 形式段落の先頭に位置するならば, その文を抽出して, Step 1. に戻る.
- 形式段落の先頭に位置しない場合, 前の文との結束性の有無を判定する. もし, 結束性がないと判定された場合は, その文を抽出する. 判定後, Step 1. に戻る.

図 1　隣接文間結束性判定による文抽出手順

以上をまとめて, 隣接文の結束性を用いて, 図 1 の手順により文抽出を行う.

このとき隣接文の結束性の有無判定を識別する必要がある. 本稿では, Support Vector Machine(以下 SVM と略記する) を用いて判定を行うことを試みた.

2.2 結束性の定義

先に述べたように, 結束性とは, 談話の整合性に大きく影響する言語的つながり [4] である.

一般に自然言語処理においては, 同じ語やシソーラスの上位で同じ分類になる語によるつながりである語彙的結束性 [3] を用いることが多い. しかし, 語彙的結束性に関しては結束性の有無の捉え方に関して個人差が生じる可能性がある. また, 同一の評定者による判定でも一貫性が取れない可能性がある. この場合, 教師あり学習のデータとして用いることが難しくなる. そこで本稿ではできるだけ語彙的結束性を具体化した形で結束性の要素として含めた.

具体的には本稿での結束性とは以下のような言語表現 (結束表現) があるものと定義する.

指示表現 (名詞的表現)
固有名詞, 普通名詞句,「こそあ」の指示名詞句, ゼロ代名詞などによる指示表現によるものを指す.

用言による結束
用言が表層として同じであることが理由で 2 つの文が結束していると判断できるものを指す.

手がかり語句に基づくもの
2 つの文がつながっていると判断できる語句 (手がかり語句) があるものを指す. 例えば以下の文[*5]では, 2 文目に接続表現「それでも」があることにより, 前の文の内容から予想されることに反する内容を述べていると判断することができる.

1. ことし落第ときまった。
2. それでも 試験は受けるのである。

2.3 隣接文間の結束性判定

本稿では隣接文間の結束性判定を SVM を用いて行う. 使用した素性は, 以下の 11 個である. なお, 以下, 読み込んだ文を S_j, S_j の直前の文を S_i, 素性の番号を k と表記する.

S_i と S_j の類似度 ($k = 1$)
固有名詞や普通名詞句の指示表現, 用言による結束においては S_i と S_j に共通の語が存在する可能性が高い. そこで素性として以下の式で表される文間類似度 $sim(T(S_i), T(S_j))$ を採用した.

$$sim(T(S_i), T(S_j)) = \frac{|T(S_i) \cap T(S_j)|}{\sqrt{|T(S_i)||T(S_j)|}} \quad (1)$$

$T(S_i)$: 文 S_i における名詞, 動詞, 形容詞の語集合

S_j の主語, 主題の省略 ($k = 2$)
以下の文[*6]のように S_j に主語, 主題がない場合, S_i の主題と同一である傾向があることがセンタリング理論による省略解消 [5] の知見により知られている. そこで, S_j に主語を表すガ格と主題をあらわす助詞「は」「も」がない場合, この素性の値を 1 とする. 主語もしくは主題がある場合はこの素性の値を 0 とする.

1. メロスは、村の牧人である。
2. 笛を吹き、羊と遊んで暮して来た。

S_j に出現する接続表現 ($k = 3..9$)
接続表現の手がかり語句による結束性を判定するために, 接続表現を以下のように 7 つに類型化したものを素性として採用している. 接続表現の類型化については [6], [7] を参考にした. なお, 実際に使用した接続表現は形態素区切りで品詞情報を含んでいる.

- 順接型 (例, だから, ですから, すると)
- 逆接型 (例, しかし, ですけども, だが)
- 添加型 (例, それから, さらに, しかも)
- 対比型 (例, 一方, 逆に, というより)
- 同列型 (例, つまるところ, ようするに)
- 転換型 (例, ところで, さて, ともあれ)
- 補足型 (例, というのは, ただ, ちなみに)

これらの接続表現が文中に出現した場合, 各類

[*5] 太宰治 『逆行』

[*6] 太宰治 『走れメロス』

表 1　作品 1 における判定結果

		被験者 1	
		有	無
被験者 2	有	22	4
	無	5	21

表 2　作品 2 における判定結果

		被験者 1	
		有	無
被験者 2	有	35	7
	無	13	6

表 3　作品 3 における判定結果

		被験者 1	
		有	無
被験者 2	有	13	2
	無	8	4

型の素性の値を以下のように定める．

$$\max_{\forall t \in Conj} \{1/(1 + C_{Num}(t))\} \qquad (2)$$

$Conj$: 文中に出現する接続表現の集合
$C_{Num}(t)$: 接続表現 t の属する文節番号
(ただし，文頭の文節番号を 0 とする)

つまり，接続表現が文頭に近いところに出現するほど素性の値は大きくなる．これは，以下の文[*7]のように複文の場合，前の節との結束を考慮しなければならないためである．

1. 下人は、大きな嚔(くさめ)をして、それから、大儀(たいぎ)そうに立上った。

S_j に出現する照応表現 ($k = 9..11$)

日本語文章で結束性の一つである指示詞，代名詞の先行詞の先行詞が一つ前の文もしくは同一文に多く出現するという調査結果 [8] に基づき，指示表現を三つの系統にまとめたものを素性として採用した．この系統立てについては [6]，[9] を参考にした．なお，実際に使用した接続表現は形態素区切りで品詞情報を含んでいる．

- コ系統 (例，この，このような，こいつ)
- ソ系統 (例，その，そのような，そいつ)
- ア系統 (例，あの，あのような，あいつ)

これらの照応表現は文中に出現した場合，接続表現と同様の重み付けを行う．これは複文の場合，文内照応を考慮しなければならないためである．

3　実験

3.1　内容

理工系の大学生 2 名に対して，2.2 節で述べた結束性の定義にしたがって短編三作品の隣接文間結束性の有無判定を行ってもらい，その結果をもとに以下のことを調べた．

1. 結束性の定義によりどのくらい個人差なく判定することができるか．
2. SVM により結束性の有無判定をどのくらいの割合で識別できるか．

今回，形態素解析器に ChaSen[*8]，構文解析器に CaboCha[*9]を使用した．また，学習ツールに SVM$^{\text{Light}}$[*10]を使用し，結束性のあるものを正例，結束性のないものを負例として学習させた．このとき，負例に対する正例の訓練誤差によるコストファクタを (負例の数)/(正例の数) と設定した．

実験には以下の作品を使用した．

作品 1. 宮本百合子『雨の昼』
作品 2. 田中貫太郎『阿芳の怨霊』
作品 3. 夏目漱石『子規の画』

3.2　個人差の一致度に関する実験結果

作品 1 から 3 における被験者の隣接文間結束性判定結果を表 1 から 3 に示す．また，被験者間の判断の一致度を表す尺度である κ 統計量は表 4 のようになった．

表 4　各作品における κ 統計量

作品	κ 統計量
作品 1	0.65
作品 2	0.16
作品 3	0.21

3.3　SVM による識別結果

各被験者の作品 1 から 3 におけるクローズドテストの結果を表 5 に示す．なお，クローズドテストは Leave-one-out と交差検定法で行っている．表の要素は，(再現率)/(精度) の順で表記している．

また，各作品に対して他の 2 作品を学習させたモデルにおける識別結果を表 6 に示す．

[*7] 芥川龍之介 『羅生門』

[*8] http://chasen.naist.jp/hiki/ChaSen/
[*9] http://chasen.org/ taku/software/cabocha/
[*10] http://svmlight.joachims.org/

表 5　SVM によるクローズドテストの識別結果

	作品 1	作品 2	作品 3
被験者 1	44.00/61.11	52.08/83.33	52.38 / 68.75
被験者 2	62.50/78.95	62.50/78.12	60.00 / 60.00

表 6　各作品を他の 2 作品で学習させたモデルにおける識別結果

	作品 1	作品 2	作品 3
被験者 1	50.00/32.00	75.00/25.00	91.67/52.38
被験者 2	44.50/33.00	72.34/85.00	66.67/53.33

4　考察

κ 統計量に基づく個人差の一致度に関する実験から作品によっては個人差が多く見られることがわかった．このことから，個人差のより少ない結束性を規定する必要があると考えられる．

SVM の識別によるクローズドテストの結果から，個々の学習データが十分識別に反映されているとはいえないことがわかった．これは結束性判定に寄与する素性が十分確保できなかったことや，類語を考慮していないことが原因であると考えられる．また，各作品を他の 2 作品で学習させたモデルにおける識別結果から，作品によって識別結果が大きく異なることがわかった．このことから，例えば，作品に出現する語などでクラスタリングを行い，そこから各作品に応じた識別器や素性を検討する必要があると思われる．

5　まとめ

本稿では小説の指示的要約を対象とした隣接文間の結束性の有無判定による文抽出手法について提案した．また，人手による隣接文間の結束性の有無判定に基づいて SVM によってどれぐらい識別できるかについて実験を行った．

今後の課題として，識別に用いるための素性についての検討や，隣接文間の結束性の有無判定による文抽出要約がどのような作品で適応できるかについて調査を行う必要があると思われる．

謝辞

実験用文章，本稿引用文として青空文庫の作品を使用した．青空文庫に携わっているボランティアの方々に感謝いたします．

また，本研究の一部は文部科学省 21 世紀 COE プログラム「インテリジェントヒューマンセンシング」および文部科学省科学研究費特定領域研究 (B)(2)16092213 の援助により行なわれた。

参考文献

[1] 『人文系研究のための言語データ処理入門』，朝倉日本語新講座６　朝倉書店，1983.

[2] 山本和英, 増山繁, 内藤昭三: “文章内構造を複合的に利用した論説文要約システム GREEN”, 自然言語処理, Vol.2, No.1, pp.39-55, 1995.

[3] Michael Halliday and Ruqaiya Hasan: “Cohesion in English”, Longman, 1976.

[4] 田窪行則, 西山佑司, 三藤博, 亀山恵, 片桐恭弘: 『談話と文脈』, 言語の科学７　岩波書店, 2004.

[5] Kameyama, M.: “A Property-Sharing Constraint in Centering”, Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics, pp.200-206 (1986).

[6] 望月源, “語彙的連鎖を用いたパッセージ抽出とその応用に関する研究”, 北陸先端科学技術大学院大学，情報科学研究科博士論文，1999.

[7] 市川考: “国語教育のための文章論概論”, 教育出版, 1978.

[8] 藤澤伸二，増山繁，内藤昭三: “日本語文章における照応・省略現象の基本的検討”, 情報処理学会論文誌 Vol.34, No.9, pp.1909-1918, 1993.

[9] 庵功雄: “新しい日本語学入門 ことばのしくみを考える”，スリーエーネットワーク，2001.