自然言語処理　150－4
(2002. 7. 15)

新聞記事要約のための文間参照関係の認定と文抽出アルゴリズム

石井 弘志　古郡 廷治
電気通信大学 電気通信学研究科 情報工学専攻
〒182-8585 東京都調布市調布ヶ丘 1-5-1

E-mail: ishii-h@phaeton.cs.uec.ac.jp, furugori@cs.uec.ac.jp

あらまし 近年、自動要約の研究が盛んに行われており、その中で文抽出型の要約は多くの手法が提案されている。文抽出型の要約手法における問題点として、抽出した文に指示表現や接続表現があった場合に、それらが参照する先行表現を含む文が要約に含まれない場合がある。本稿では、新聞社説記事中に現れるコソ系の指示表現、「同・両」の指示表現、文頭表現、主語省略文について、どのように文間の参照関係を認めるかについて述べる。また、要約文に参照元の表現のみが現れないようにする文抽出アルゴリズムを示す。文間参照関係を認定するシステムの精度を測った結果、人手で作成したデータと比較し再現率約90%, 適合率約70%を得た。

キーワード 自動要約, 抄録, 文間参照関係, 文抽出

The Problem of Reference Relations and Sentence Extractions in Summarization Systems

Hiroshi ISHII　Teiji FURUGORI
Department of Computer Science, University of Electro-Communications
1-5-1 Chofugaoka, Chofu, Tokyo, 182-8585 Japan

E-mail: ishii-h@phaeton.cs.uec.ac.jp, furugori@cs.uec.ac.jp

Abstract Sentence extraction method is a practical means to devise summarization systems. An obvious weakness of this method is that it usually gives no considerations for reference relations among the elements (words) in the sentences to be extracted. In this paper, we try to get around the weakness for the sentences that contain the determiners starting with *ko, so, dou,* and *ryou*, zero pronouns, and conjunctions are involved. We then offer an algorithm and an experimental result for extracting sentences that alleviates the reference problem.

Keyword automatic summarization, extract, reference relations, sentence extraction

1. はじめに

近年、自動要約の研究が盛んに行われており、その中で文抽出型の要約は文書の領域にそれほど依存せず、また処理が複雑とならない手法として、多くの手法が提案されている[1]。

文抽出型の要約手法では、それぞれの文の重要度を計算し、重要な文を抜き出したものを要約とする。

この手法の要約作成における問題点として、抽出した文に指示表現や接続表現があった場合に、それらが参照する先行表現を含む文が要約に含まれない場合がある。このような要約は一般に解釈が容易でないため、文章として読みづらく、また誤った解釈で読んでしまう場合もある。そのため、指示表現や接続表現を適切に処理しないと、要約としての適切性を失わせる要因となる。

本研究の目的は、指示表現や接続表現について、文間の参照関係を認定する手法を提案し、その結果を用い、できる限り受容可能性（acceptability）の高い要約を作成することである。

文間の関係を調査した研究や、自動要約における結束性の処理を扱った研究がいくつか報告されている。柴田ら[2]は、新聞社説記事においてどのような照応現象があるかを調査している。山本ら[3]は要約採用文の文頭に指示詞や接続詞などの手がかり語が出現する場合や、主語が省略されている場合には、前文を採用することで近似的な結束性の処理を行っている。また、難波ら[4]は抄録の読みにくさの要因とそれを解消する方法を調査し、そのうちの一部について、主に要素の削除による書き換えルールを実装している。大塚ら[5]は連体詞形態指示詞の先行詞が重要文に含まれていない場合にその先行詞の補完を行う手法を提案している。

本稿では、新聞社説記事中に現れるコソ系の指示表現、「同・両」の指示表現、文頭表現、主語省略文について、どのように文間の参照関係を認めるかについて述べる。特に、コソ系の指示表現、同・両の指示表現については、文頭・文中に関わらず一貫した処理を行う。文間参照関係の認定手法を示したあと、参照関係を保つ要約を作成する文抽出アルゴリズムを示す。

2. 参照関係の認定

この章では、新聞社説記事中の指示表現や接続表現について、どのように文間の参照関係を認めるかを述べる。本稿で扱う参照関係は、次の4種類である。

- コソ系の指示表現
- 「同・両」の指示表現
- 文頭表現
- 主語の省略

本手法で参照関係を認定する手順は次の通りである。

(1) 原文章を形態素解析システムjuman[6]で形態素解析する。

(2) すべての文に関して上記に述べた各種の参照関係を調べる。

(3) Rという関係でi文目がj文目を参照しているという関係を、Ref[i,j,R]という形式で表す。

参照関係の作成の仕方には、次の3種類を定義する。

1文参照 i文目からj文目への参照(Ref[i,j,R])とする（図1）。

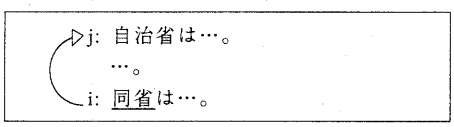

図1　1文参照

範囲参照 i文目からn文前の文までを全て参照(Ref[i,i-1,R], Ref[i,i-2,R], …Ref[i,i-n,R])とする（図2）。

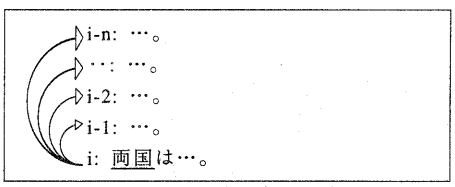

図2　範囲参照

前段落・前文参照 該当参照表現の出現する文が段落先頭の文の場合、一段落前の文を全て参照(Ref[i,i-1,R], Ref[i,i-2,R], …Ref[i,一段落前の先頭の文,R])とする。指示表現の出現する文が段落先頭ではない場合、1文前の文を参照(Ref[i,i-1,R])とする（図3）。

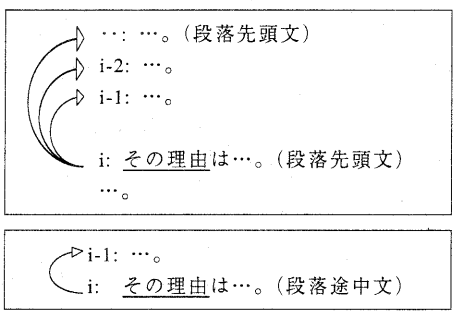

図3　前段落・前文参照

以下、それぞれの参照関係をどのように認定するかを述べる。

2.1. コソ系の指示表現

「これ」や「そのN」などの、コソ系の指示表現では、指示する対象によって次の3種に分類する。

A. 語・句を指示すると推測できる表現
例：その地震　この支部
B. 節・文を指示すると推測できる表現
例：その理由　これにより…
C. 指示先のない表現
例：これまで　ここ数日

分類の方法は、次のように行う。

(1) “これまで”、“これから”、“このところ”、“[指示詞]+数日”のような、指示先のない表現のリストをあらかじめ作成し、これに該当する表現はCとする。
(2) “[指示詞]+[名詞または未定義語以外の形態素]”はBとする。また、“[指示詞]+こと”や“[指示詞]+理由”のような節・文を指示すると推測できる表現のリストをあらかじめ作成し、これに該当する表現はBとする。
(3) (1),(2)以外の“[指示詞]+[名詞または未定義語]”の表現はAとする。

上記A, B別に、文間の参照関係を次のように認定する（R=コソ，指示表現が出現する文をi文目とする）。ただしA, Bともに、指示表現の直前が“および or と”となっている場合、同一文内参照として参照関係は作らない。

（例1）…政党**および**その政治資金団体…。
（例2）…天気**と**その変化…。

A. 語・句を指示すると推測できる表現

指示詞に後続する名詞、未定義語、接尾辞を連結した文字列を、出現する箇所から前方方向に検索し、部分一致する文字列がj文目に見つかれば、i文目からj文目への1文参照（Ref[i,j,コソ]）とする。同一文内に部分一致文字列が見つかる場合には参照関係は作らない。但し、検索された語に、コソ系の指示表現が前置されている場合、さらに前方を検索する。

指示詞に後続する文字列が見つからない場合、i文目からn文前の文までの範囲参照（Ref[i,i-1,コソ], Ref[i,i-2,コソ], …Ref[i,i-n,コソ]）とする。

（例1）…選挙区を単位とする**支部**を…。…。…これら**支部**…。（2文前を参照）
（例2）…高層ビル群や大規模な地下街が…。高速道路も…。そうした構造物は…。（n文前までを範囲参照）

B. 節・文を指示すると推測できる表現

当該指示表現のある文の、文頭から指示表現までに次のパターンのいずれかがマッチする場合、同一文内参照として参照関係は作らない（パターン中の“*”は0個以上の出現を表す）。

Pattern1 [動詞 or 助動詞]+[接尾辞]*+[助詞]*+読点
（例）…ある日突然、社長交代が**行われ、**そのきっかけは…
Pattern2 [判定詞]+[助詞]*+[接尾辞]*+[助詞]*+読点
（例）…原告の上告は差し止め請求が中心**だったが、**この点が棄却されたのは…

以上のパターンにマッチしない場合、前段落・前文参照とする。すなわち、指示表現の出現する文が段落先頭の文の場合、一段落前の文を全て参照（Ref[i,i-1,コソ], Ref[i,i-2,コソ], …Ref[i,一段落前の先頭の文,コソ]）とする。指示表現の出現する文が段落先頭ではない場合、1文前の文を参照（Ref[i,i-1,コソ]）とする。

（例1）　…日本の株主は…。…個人投資家の発言力は…。
　　…こうした傾向は…。（前段落参照）
（例2）…明治の京都人に…。…市民が誇るのも、ここに由来する。（前文参照）

例外として、コ系の指示表現で、次の文が引用表現（“「” “」”で囲われている文）の場合、i文目からi+1文目への1文参照（Ref[i,i+1,コソ]）とする。
（例）…はこう述べている。「……。」

2.2.「同・両」の指示表現

「同…」や「両…」などの、「同・両」の指示表現についても、コソ系の指示表現と同様に、指示する対象によって次の3種に分類する。

A. 語、句を指示すると推測できる表現
例：同省　両国
B. 節・文を指示すると推測できる表現
例：同様　同時
C. 指示先のない表現
例：同僚　両生類

分類の方法は、次のように行う。

(1) “同僚”、“同盟”、“両生類”、“両手”のような、指示先のない表現のリストをあらかじめ作成し、これに該当する表現はCとする。
(2) “同様”や“同時”のような節・文を指示すると推測できる表現のリストをあらかじめ作成し、これに該当する表現はBとする。
(3) (1),(2)以外の“同…”または“両…”の表現はAとする。

上記A, B別に、文間の参照関係を次のように認定する（R=同・両, 指示表現が出現する文をi文目とする）。ただし「同…」では、A, Bともに、指示表現の直前が“および or と”となっている場合、同一文内参照として参照関係は作らない。

A. 語・句を指示すると推測できる表現

「同…」の場合、“同”に後続する名詞、未定義語、接尾辞を連結した文字列を、出現する箇所から前方方向に検索し、部分一致する文字列がj文目に見つかれば、i文目からj文目への1文参照（Ref[i,j,同・両]）とする。同一文内に部分一致文字列が見つかる場合には参照関係は作らない。但し、検索された語に、“同”が前置されている場合、さらに前方を検索する。

指示詞に後続する文字列が見つからない場合、i文目からn文前の文までの範囲参照（Ref[i,i-1,同・両], Ref[i,i-2,同・両], …Ref[i,i-n,同・両]）とする。

（例1）…自治省は…。…。…。…同省は…。（3文前を参照）
（例2）…クリントン米政権との…。…。…同大統領の…。（n文前までを範囲参照）

「両…」の場合、「両国」で「日本」「アメリカ」を参照する場合のように、「両」のうしろの文字列が先行表現に部分一致しない。そのため、n文前までの文を全て参照とする。但し、「N1、N2両…」「N1とN2の両…」といった表現は同一文内参照として参照関係は作らない。

B. 節・文を指示すると推測できる表現

当該指示表現のある文の、文頭から指示表現までに、2.1節で示したPattern1、2のいずれかのパターンがマッチする場合、同一文内参照として参照関係は作らない。パターンにマッチしない場合、前段落・前文参照とする。

（例1）…賃上げ純増ゼロをちらつかせている。…対立する気配である。…一段と厳しくなるだろう。
…昨年の春闘も同様だった。（前段落参照）
（例2）…が掲載されていた。同時に…。（前文参照）

2.3. 文頭表現

「そして」や「しかし」などの前文との接続的な表現が文頭に出現する場合、前段落・前文参照を行う（R=文頭）。

接続的な文頭表現の認定は次のいずれかに該当するものとする。

①jumanで形態素解析した際に、文の先頭の形態素の品詞が[接続詞]となっているもの
②[接続詞]以外で接続的表現と認められる表現（表1に例を示す。）

表1 [接続詞]以外の接続的表現の例

また	さらに	が、	特に	とくに	ただ、
例えば	たとえば	一方	その上で	そのうえで	
むしろ	とりわけ	あえて言えば	あえていえば		
いずれも	どうも、	と書くと	中には	なかには	
中でも	なかでも	現に	まして	要は	つまり
にもかかわらず	以上	この際、	このさい、		
と言うのは	というのは	次は			

2.4. 主語の省略

一般に、文の主語が抜けている場合、文の理解が困難となる。省略されている文の主語は多くの場合前文に出現することが多いので、主語の省略が起こっている文では前段落・前文参照を行う(R=省略)。主語の省略の認定は、助詞の「が」「は」「も」が存在しない文に主語の省略が起きていると判定する。

（例）…。献金問題である。（前文参照）

しかし、文の種類によっては、「私たち」が隠れた主語になっていて、その文のみで理解が問題ない場合がある。「…を警戒しよう。」というような文がそれである。このような、「…しよう。」という意志を表す表現が述語の文は「私たち」が省略されていると考えられ、前文を補う必要性は薄いと思われる。そのため、参照関係は作成しない。述語が、意志を示す表現であるかどうかは、jumanで形態素解析した際に、述語動詞が「意志形」の動詞であることによって判断を行う。

そのほか、「首相の早期訪中に期待したい。」のような、接尾辞の“たい”で終わる文や、「大きな話題といえば、自動車業界の見事な再生ぶりだ。」のような“といえば+(…)+判定詞”という文についても参照関係を作成しない。

3. 参照関係を保存する文選択のアルゴリズム

文選択型の要約で、2 節で述べた参照関係を保存する文選択アルゴリズムを示す。これは、要約に於いて参照元のみの出現が起こらないようにするアルゴリズムである。従って、参照元と参照先のペア、または参照先のみの出現を許すアルゴリズムとなっている。

はじめに、文の加工をせずに文をそのまま抽出する場合のアルゴリズムについて述べ、次に、文の加工を許すことで必ずしも参照先の文が出現しなくてもよい場合の文抽出方法を示す。

3.1. 文を加工しないで抽出する場合

(前処理) 図 4 にあるように相互に参照関係がある 2 文について、まとめて 1 文のようにみなす。相互に参照関係がなくなるまで、まとめる処理を繰り返す。

(1) 任意の重要度順序に基づいて文の採用順を決める。
(2) 重要度の高い文から 1 文をとる（S とする）。
(3) S が参照元である参照関係 Ref[S,x,R]について、全ての参照先の文が採用済みかどうか調べる。
(4) 全ての参照先の文が採用済みであるか、S が参照元である参照関係が存在しない場合、S を採用する。参照先の文が 1 つでも採用されていなければ、S を採用候補プールに入れる。
(5) 採用を行う操作を行うごとに、そのとき採用した文が要約に含まれることによって、採用候補プールに格納されている文が採用可能になったかどうかを調べ（参照先の文が採用済みかどうか）、可能ならば採用候補プールから文を取り出し採用する。
　採用時に要約率のチェックを行い、要約率の条件を満たす場合、文選択処理を終了する。
(6) (2)に戻る。

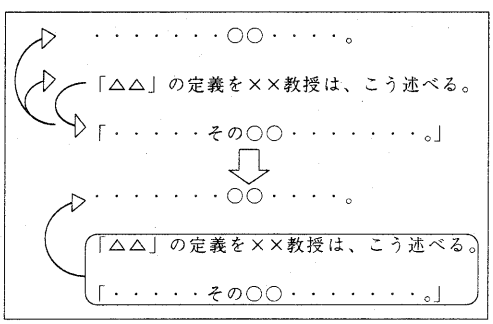

図 4　相互参照文をまとめる処理

3.2. 要約文の抽出後に文の加工を行う場合

接続詞の削除を行ったり、先行詞を補完することによって、文間の参照関係を切り離すことが可能である。

要約文の抽出をした後にこのような処理を行うことで、参照先の文が存在しなくても受容可能な要約が作成できる。ここでは、このような選択制限の緩い文抽出方法を示す。

(準備) 可能なものについて、参照先が要約に存在しなくても要約が理解可能となる加工ルールを決める。これには、文頭の接続詞を削除するルールや、難波ら[4]の抄録書き換え手法、大塚ら[5]の連体詞形態指示詞の先行詞の補完手法などを用いることができると思われる。

(1) 3.1 節と同様の文抽出アルゴリズムで文を抽出する。但し、加工ルールのない参照関係のみについて参照先が採用済みかどうか調べる。
(2) 後処理として、生成された要約に加工ルールのある参照があり、その参照先の文が採用されていなければ加工ルールを適用する。

4. 評価実験

2 節で述べた文間参照関係の認定手法がどの程度の精度で正しい参照関係となるかを調べた。正しい参照関係であるかどうかは、計算機上に実装したシステムが判定した参照関係と、人手で作成した正解データとの比較を行うことで判断した。正解データに用いたのは、新聞社説記事 20 編[7]で、1 記事の平均文数は 31.2 文である。正解データに含まれる文間参照関係のある表現は、全体でコソ系の指示表現が 123 個、「同・両」の指示表現が 15 個、文頭表現が 94 個、主語省略文が 35 個であった。正解の文間参照関係の数は、コソ系の指示表現が 167 個、「同・両」の指示表現が 16 個、文頭表現が 121 個、主語省略文が 36 個であった。表現の数より参照関係の数のほうが多くなるのは、1 つの表現で複数の文を参照していることがあるためである。

システムの正解率の判定方法には、再現率と適合率を用いた。再現率と適合率の算出式を以下に示す。

$$\text{再現率} = \frac{\text{正解である参照関係の数}}{\text{システムで認定した参照関係の数}}$$

$$\text{適合率} = \frac{\text{正解である参照関係の数}}{\text{人手で認定した参照関係の数}}$$

まず、文頭表現と主語の省略については、文頭表現が再現率 0.934, 適合率 0.883、主語の省略が再現率 0.889, 適合率 0.604 であった。

表 2 に、参照関係全体の再現率・適合率、および、コソ系指示表現と「同・両」の指示表現について、範囲参照の範囲(n 文前まで)を変化させて再現率・適合率を算出した結果を示す。

表 2　参照関係全体、コソ系指示表現、同・両の指示表現の再現率・適合率

n	参照関係全体 再現率／適合率	コソ系指示表現 再現率／適合率	同・両指示表現 再現率／適合率
1	0.885／0.741	0.862／0.706	0.750／0.571
2	0.897／0.723	0.886／0.692	0.750／0.444
3	0.900／0.700	0.886／0.664	0.813／0.394
4	0.900／0.677	0.886／0.638	0.813／0.333
5	0.903／0.657	0.886／0.614	0.875／0.311

表 2 において n=5 での参照関係全体での再現率が 0.903 にとどまっている原因は、主に前段落・前文参照で先行表現をカバーしきれていないのが原因である。

なお、システムが文間に参照関係があると判定した表現の数は、全体でコソ系の指示表現が 128 個、「同・両」の指示表現が 18 個、文頭表現が 93 個、主語省略文が 44 個であった。また、システムが出力した文間参照関係の数は、n=3 の場合で、コソ系の指示表現が 223 個、「同・両」の指示表現が 33 個、文頭表現が 128 個、主語省略文が 53 個であった。

付録に、文の重要度をランダムで与えた場合の、重要度の高い順に文を抽出した抄録と、3.1 で示したアルゴリズムで文を抽出した抄録を比較して示す。

5. 考察

提案した手法により評価実験を行った結果では参照関係全体で再現率約 90%、適合率約 70%を得ることができた。

適切な参照関係を認定できなかった場合として、次のようなものがあった。

（例 1）…。…。…。…。…そう弁明してきた。（システムは前文参照と認定。正解は前の 4 文を参照）

（例 2）…米国自動車業界は、その政治力に頼って…（システムは n 文前までの範囲参照と認定。正解は同一文内参照）

（例 3）…米中両国の…。（システムは n 文前までの範囲参照と認定。正解は同一文内参照）

（例 4）…。女性たちが国境を**超えて**、その視点を…。（システムは同一文内参照と認定。正解は前文参照）

このような表現には、ヒューリスティックスで一部対応できると考えられる。しかし、意味を捉えないと正確には判断できないものもある。また、本稿で認定を行った参照関係のほかにも、例えば「冒頭に紹介したのは…」「前者については…」「第二に…」といったような、単独では理解が困難な表現があった。

6. おわりに

本稿では、文間に存在する参照関係の認定方法と、要約文に参照元の表現のみが現れないようにする文抽出アルゴリズムを示した。システムの精度を測った実験結果から、本手法を用いることである程度の受容可能性を有する要約が作成できると思われる。

要約文を作成する際に参照関係を保つようにすることは、要約の読みやすさ、適切性を保つことにつながり、誤った解釈をすることも少なくできると考えられる。

今後の課題としては、考察で述べた表現にも適切に対応すること、文抽出をする際に重要度が最大となる文セットを適切に選択することがあげられる。

参考文献

[1] 奥村学, 難波英嗣, “テキスト自動要約に関する研究動向”, 自然言語処理, vol.6, no.6, pp.1-26, 1999.

[2] 柴田昌宏, 田中理, 福本淳一, “新聞社説記事における照応現象”, 情報処理学会第 40 回全国大会, 5F-4, pp.474-475, 1990.

[3] 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.

[4] 難波英嗣, 奥村学, “書き換えによる抄録の読みやすさの向上”, 情報処理学会研究報告, NL-133, pp.53-60, 1999.

[5] 大塚敬義, 内海彰, 奥村学, “要約文生成における照応処理”, 電子情報通信学会 信学技報, TL2001-4, pp.19-26, 2001.

[6] 黒橋禎夫, 長尾真, 日本語形態素解析システム JUMAN version 3.61, 京都大学大学院情報科学研究科, http://pine.kuee.kyoto-u.ac.jp/, 1999.

[7] CD-ROM 毎日新聞'94, 日外アソシエーツ株式会社

付録

文の重要度をランダムで与えた場合の抄録例を示す（網掛け部分は、先行表現の出現していない参照表現である）。

重要度の高い順に文を抽出した抄録

こんな元気の出るあいさつを聞きたかった。
今春のビジネス社会での大きな話題といえば、米国の自動車業界の見事な再生ぶりだ。
“甘え”といえばいまひとつ気になるのが、昨年末からジリジリと進んでいる円安傾向である。
だが、多くの企業トップは、黒字削減が実効をあげない現在、それは望むべくもない水準であることを知っている。
苦境だからといって、政府や身内の従業員に甘えることは許されない。
まして、消費者の欲求に応えられないような経営者は、失格の烙印を押されてもやむを得ないだろう。

参照関係を保存する文選択アルゴリズムによる抄録

企業という利益追求集団のリーダーは、厳しい時だから明るく振る舞い、夢を語る責務があろう。
いま、日本の企業に求められているのは、他力をあてにせぬ企業家精神の発揮なのだ。
今春のビジネス社会での大きな話題といえば、米国の自動車業界の見事な再生ぶりだ。
“甘え”といえばいまひとつ気になるのが、昨年末からジリジリと進んでいる円安傾向である。
苦境だからといって、政府や身内の従業員に甘えることは許されない。
まして、消費者の欲求に応えられないような経営者は、失格の烙印を押されてもやむを得ないだろう。