文外照応を含む文の検出による抽出型要約の品質向上

西川 仁　長谷川 隆明　松尾 義博　菊井 玄一郎
日本電信電話株式会社 NTT サイバースペース研究所
{ nishikawa.hitoshi, hasegawa.takaaki
matsuo.yoshihiro, kikui.genichiro } @lab.ntt.co.jp

概要

本稿では，重要文抽出に基づく要約を行う際に，照応詞を含む文を検出し，それらの文は先行詞を含むと思われる文と共に要約に含めることによって，要約の品質が改善されることを示す．要約の対象とする文書集合に対して文分割を行ったのち，照応詞を含む文を検出する分類器によってそれらの文を検出する．照応詞を含む文はそれらの文の前の文と連結し，一つの文として扱い，要約器への入力とする．本手法を重要文抽出の前処理として用いることで，要約に含まれる，先行詞のない照応詞やゼロ代名詞の数が減少することを実験によって示す．

1　はじめに

現在の文書要約技術の多くは文を単位にした処理を行っている [7]．一般には，まず入力された文書集合を文分割器により文集合に変換する．次に，文集合から，要約長を満たす文の組み合わせを，要約としての善し悪しを与える何らかの基準に基づき選び出すことによって要約は生成される．この場合，仮に要約の中に照応詞を含むような文が含まれていた場合，それらの先行詞を含む文が要約に含まれるという保証はない．その場合，要約の中には先行詞を持たない文が含まれることになり，原文書におけるその文本来の意味が曲解されたり，あるいは全く意味の通らない要約が生成される恐れがある．また，そのような要約は，要約に対する読み手の信頼を著しく損ない，要約技術の実用上の価値を大きく損なう恐れもある．

問題となる要約の例を以下に示す．

(1) この機能はとてもバッテリを消費するので，注意が必要です．
(2) 値段は少々張りますが，トータルで見ると大変お買い得なカメラです．
(3) とても面白い機能です．

以上の3つの文で1つの要約が構成されていたとき，文 (1) の「この機能」の指示するものや，文 (3) の主語がわからないことから，この要約は有用なものとは言い難い．

このような場合に対する一つの解は，入力文書集合に対して事前に照応解析 [9] を実施することである．ゼロ代名詞が含まれる文は省略された要素を事前に補っておき，照応詞は指示する対象に適切に置換することによって，以上の問題を回避できる．しかし，既存の照応解析の精度は，その問題の難しさから，必ずしも十分ではない．特に省略された要素を同定することは容易ではないため，重要文抽出の前処理として用いることは必ずしも得策ではない．

そこで，本稿では，先行詞の同定は行わず，単に照応詞やゼロ代名詞を持つと思われる文を検出し，それらの文は前の文と連結させ一つの文として扱うことによって，この問題を部分的に回避することを提案する．多くの場合，先行詞は直前の文に存在することが知られているため，そのような文は前の文と連結させて単一の文として扱うことによって，要約中に先行詞が存在しない照応詞やゼロ代名詞が含まれる数を減少させることができると期待される．

このような要約の前処理は文分割同様高速になされる必要がある．そこで，本稿では，文に対して係り受け解析などは適用せず，形態素解析のみを適用した結果を特徴量として用いて，確率的な分類器を構築し，問題となる文を検出する．

以下，2節では関連研究について述べる．3節では要約に含まれる照応詞を分析し，その結果について述べる．4節では問題となる照応詞を含む文を検出する分類器の構築について述べる．5節では評価実験の設定について述べ，6節では実験の結果，要約に含まれる先行詞のない照応詞の数が減少することを示す．7節では本稿についてまとめる．

2　関連研究

先行詞のない照応詞が要約の問題となることは以前から知られており，規則によって照応詞を削除するなどの方法が取られている [4]．また，照応詞を含む文は，その前の文と連結させることも提案されている [8]．一方で，問題となる照応詞を網羅的に検出，あるいは削除できる規則の集合は明らかではない．

他のアプローチとして，文書を段落に分割する方法 [1] がある．文書を適当な単位に分割することで，照応の問題が生じがたい単位を得られることが期待できる．しかし，照応関係が段落境界を跨らない保証はなく，また要約を構成する単位が文に比べて長くなって

しまうため，柔軟に要約を生成できない恐れがある．

3 分析

照応解析および照応関係の付与に関する先行研究 [10, 9, 2] を参考に，要約の品質を阻害する照応関係について分析する．

3.1 コーパス

まず，要約に含まれる省略や指示表現のうち、要約の内容の理解を妨げるものを分析する．本稿ではレビューを要約の対象とする．ウェブ上から，PC と PC 周辺機器，PC パーツ，AV 機器，生活家電の 4 つのドメインのレビューを収集した．収集したレビューのうちから，それぞれのドメインについてランダムに 25 種類の商品を選び，それらに付随する 1 つ以上のレビューを入力として合計 100 種類の要約を生成した．要約長は 240 文字とした．要約器には重要文抽出に基づく要約器 [5] を用い，要約器が必要とするパラメータは要約を行う文書と同一のドメインのデータから推定した．

3.2 問題となる場合

文外照応　文中に現れた照応詞が文内照応である場合，その文は単独で要約に用いることができるため，先行文脈を与える措置は不要となる．一方，文外照応である場合は文脈を与える必要がある．そのため，文内照応と文外照応を区別する必要がある．

(4) **ドライバー**をダウンロードし，**それ**をインストールすれば OK.

(5) しかし，今では誰もが携帯電話を持っているため，**その役目**は終わりつつあります．

特に，複数の節からなる文の後方の節に照応詞がある場合，文内照応と文外照応の区別が難しい．文 (4) の「それ」の指示するものは「ドライバー」であるため問題は生じない．一方，文 (5) には先行文脈が不可欠である．

また，文外照応と外界照応を区別する必要がある．

(6) **このモニタ**は独特の機構によって自在に画面が上下する．

文 (6) の文頭の「このモニタ」が示すものは，レビューが記述の対象としているモニタであり，外界照応となっている．要約を読み手に提示する場合に，同時に記述の対象も提示される場合は，このような外界照応は問題とならない．本稿では外界照応は要約の問題にならないと仮定する．そのため，文外照応と外界照応を区別する必要が生じる．

(7) 指紋認証の**機能**がついています．

(8) が，**この機能**はあくまでオマケと思います．

一方，文 (8) は文書内照応であり，要約の問題となる．

指示詞を伴わない照応　指示詞を伴わない場合，まず文に含まれる名詞句が要約の問題となるものであるかを判断する必要がある．

(9) 問題はパネルの**質**．

(10) 同価格帯のものと比べ**質**が悪い．

文 (10) のみでは,「質」の指示するものを同定できず，文意を正しく理解できない．この例は「この」「その」といった指示連体詞を伴わないため，検出が難しい．

また，省略された要素（ゼロ照応）の検出も大きな問題となる．

(11) **画面**の質が低いと思う．

(12) あまり綺麗に（φ が）見えない．

普通名詞句の照応の場合と同様に，指示詞を手がかりとすることができないため，別のアプローチが必要となる．

3.3 分布

問題となる照応の分布を表 4 の「本手法なし」の列に示す[1]．代名詞照応とゼロ照応の数は同程度，名詞句照応の数がそれらより若干少なくなっている．

4 検出モデル

上述した，問題となる照応関係は，照応解析器，述語項構造解析等を利用すればある程度は検出が可能である．一方，一般にそれらの深い解析を行う解析器は係り受け解析の結果を特徴量として用い解析を行う．本稿では一種のフィルタとして本手法を用いるため，形態素解析までの浅い解析で得られる結果を特徴量として用いる確率的な分類器を構築する．また，先行詞の同定までは試みずに，あくまで個別の文に対し照応や省略の有無を検出することを試みる．

4.1 ロジスティック回帰モデル

ロジスティック回帰モデルを用いて，入力された文書を構成する文全てに対し，逐次的に分類を実施し，問題となる文を検出する．

$$\hat{y} = \underset{y \in \{0,1\}}{\operatorname{argmax}}\, p(y|\mathbf{f}(x); \mathbf{w}), \tag{1}$$

$$p(y = 1|\mathbf{f}(x); \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w} \cdot \mathbf{f}(x))}. \tag{2}$$

ここで，$\mathbf{f}(x)$ は文 x の特徴ベクトル，$\mathbf{w}$ は重みベクトル，$p(y = 0|\mathbf{f}(x); \mathbf{w}) = 1 - p(y = 1|\mathbf{f}(x); \mathbf{w})$ である[2]．

[1] 間接照応は名詞句照応に含めた．

[2] すなわち，以下の推論を，$\theta = 0.5$ として暗に仮定している．

$$\hat{y} = \begin{cases} 1 & (p(y|\mathbf{f}(x); \mathbf{w}) \geq \theta) \\ 0 & (otherwise) \end{cases}.$$

4.2 特徴量

以下に述べる特徴量から特徴ベクトル $\mathbf{f}(x)$ を構成する．

格助詞の有無　ガ格，ヲ格，ニ格などが省略されていることを検出するため，それぞれの格助詞の有無を特徴量として用いる．

指示詞の有無　「この」「その」といった指示詞の有無を特徴量として用いる．

格助詞，指示詞の文書中での位置　それぞれの格助詞，指示詞の文書中での位置，すなわち先頭からの文数を特徴量として用いる．文書の先頭の文が問題となる照応詞を持つ可能性は低い．

格助詞，指示詞の文中での位置　それぞれの格助詞，指示詞の文中での位置，すなわち先頭からの単語数を特徴量として用いる．文 (4) に示すように，文の後方にある指示詞は問題とならない可能性が高い．

格助詞，指示詞の前にある読点の数　それぞれの格助詞，指示詞の前にある読点の数を特徴量として用いる．文 (4) のように，前に読点がいくつもある指示詞は問題とならない可能性が高い．

前の読点からの距離　それぞれの格助詞，指示詞より前にある読点のうち最も近いものからの距離（単語数）を特徴量として用いる．前に読点がないものは文頭からの距離を代わりに用いる．

前の文の単語数との差　前の文と単語数が大幅に異なる場合，特に前の文と比べ単語数が大幅に減少している場合何らかの省略が行われている可能性が高い．

前の文と同一の名詞の数　文 (9) および (10) に示されるように，前の文に同一の名詞がある存在する場合，そこに照応関係が存在し，さらに要約にとって問題となる可能性がある．

前の文とのコサイン類似度　上の特徴量と同様の意図により，前の文とのコサイン類似度を特徴量として用いる．

5 評価実験

5.1 コーパス

3 節と同様に，ウェブ上から，PC と PC 周辺機器，PC パーツ，AV 機器，生活家電の 4 つのドメインのレビューを収集し，それぞれのドメインについてランダムに 25 種類の商品を選択した．それらの商品に付随する 1 つ以上のレビューを文に分割したのち，各文に対し前の文との連結が必要であるか否かを示すラベルを付与した．各ドメインの事例数（すなわち文の総数）を表 1 に示す．括弧の中の数字はそれぞれ正例，負例の数である．正例は問題となる照応詞を持つ文の数である．要約器も同様に重要文抽出に基づく要約器 [5] を用いた．

表 1: 各ドメインの事例数

	事例数（正例/負例）
PC と PC 周辺機器	968 (198/770)
PC パーツ	883 (106/777)
AV 機器	704 (103/601)
生活家電	711 (116/595)
全体	3266 (523/2743)

表 2: ラベル推定の適合率，再現率，F 値

	適合率	再現率	F 値
PC と PC 周辺機器	0.775	0.812	0.793
PC パーツ	0.830	0.898	0.862
AV 機器	0.848	0.876	0.861
生活家電	0.714	0.883	0.784
全体	0.792	0.865	0.824

5.2 パラメータ推定

重みベクトル $\mathbf{w}$ は確率的勾配降下法を用いて推定した．L2 正則化を加え，正則化係数は 0.01，イタレーション数は 30 回とした．

6 実験結果と考察

6.1 ラベル推定の適合率，再現率，F 値

推定されたラベルの F 値を表 2 に示す．値は，5 分割交差検定を行い，各セットに対する平均を取ったものである．

偽陽性の誤りの多くは，外界照応に起因するものである．省略された要素や，指示詞によって指示されている要素が外界照応である場合，今回の場合読み手に取って問題にはならない．しかし，その区別ができず誤って問題となる文として検出された場合が多かった．偽陰性の誤りの多くは省略の検出の失敗によるものである．目的語などが省略されていることを検出することができず，問題となる文を検出できない場合が多かった．

6.2 要約に含まれる先行詞のない照応詞の数

本手法を前処理として用いずに要約を生成した場合と，前処理として用いて要約を生成した場合で，要約に含まれる先行詞のない照応詞の数を比較した．入力は 3.1 節にて述べたものと同じものである．すなわち各ドメインにつき 25 種類の要約を生成した．入力は学習事例の中に含まれない．結果を表 3 に示す．

PC パーツのドメインでは本手法による悪化が発生

表 3: 要約に含まれる先行詞のない照応詞の数

	本手法なし	本手法あり
PC と PC 周辺機器	14	10
PC パーツ	10	11
AV 機器	12	11
生活家電	7	5
全体	43	37

表 4: 問題となる照応詞の種類の分布

	本手法なし	本手法あり
代名詞照応	17	11
名詞句照応	10	8
ゼロ照応	16	18
合計	43	37

しているが，他のドメインでは先行詞のない照応詞の数が減少した．しかし，全体の値に対してウィルコクソンの符号付順位和検定を実施したところ，有意確率は 0.252 であった．そのため，本手法により有意に先行詞のない照応詞が減少したと言い切ることはできない．

本手法を前処理として用いた場合の，問題となる照応詞の種類の分布を表 4 に示す．代名詞照応の数は大きく減少した．これは指示詞などの表層手がかりにより，比較的容易に問題となる文を検出できるためと考えられる．一方，手がかりの少ない名詞句照応，ゼロ照応に対しては，本手法が有効に働いたとは言えない．

6.3 生成された要約の内容性

本手法を用いると，要約として本来選ばれない文が，先行文脈として要約に含まれることになる．そのため，先行詞のない照応詞が減少することによる可読性向上の代償として，要約の品質，特に内容性が低下する恐れがある．そのためここでは要約の自動評価尺度である ROUGE[3] を用いて要約の内容性を評価する．

商品レビューに対する要約 [6] の際に用いた原文書集合および対応する参照要約の中から，上述の商品カテゴリに該当するものを選び出し，本手法を用いた場合と用いない場合それぞれの ROUGE-1 スコアを評価した[3]．結果を表 5 に示す．カテゴリ名の後の数字は要約の対象とした文書集合の数である．

PC と PC 周辺機器のドメインを除き，いずれのドメインでも ROUGE-1 スコアは低下した．上述したように，本来要約に含まれないはずの文が本手法により要約中に含まれることによるものであると考えられる．一方，全体の値に対してウィルコクソンの符号付順位和検定を実施したところ，有意確率は 0.197 であり，有意に ROUGE-1 スコアが低下したとは言えない．

[3]名詞，動詞，形容詞，未知語のみを用いてスコアを計算した．

表 5: 生成された要約の ROUGE-1

	本手法なし	本手法あり
PC と PC 周辺機器 (11)	0.351	0.366
AV 機器 (30)	0.340	0.312
生活家電 (17)	0.313	0.299
全体 (58)	0.334	0.318

7 おわりに

本稿では，要約に含まれた際に問題となる照応詞を含む文をロジスティック回帰モデルに基づく分類器を用いて検出し，それらの文は直前の文と同時に要約に含めることで，要約に含まれる先行詞のない照応詞の数が減少することを示した．冒頭に述べたように，先行詞を欠く照応詞は要約の全体的な品質を大きく損なうものと考えられる．そのため，ROUGE の値を多少犠牲にしてでも，照応詞に関する問題を改善することは要約の全体的な品質向上にとっては重要かもしれない．要約の，内容的品質と言語的品質の調和は，興味深い問題である．

参考文献

[1] Marti A. Hearst. TextTiling: segmenting text into multi-paragraph subtopic passages. Computational Linguistics. 23(1):33–64, 1997.

[2] 飯田龍, 小町守, 井之上直也, 乾健太郎, 松本裕治. 述語項構造と照応関係のアノテーション. 自然言語処理, 17(2):25–50, 2010.

[3] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In Proc. of the Workshop on Text Summarization Branches Out, 2004.

[4] Hidetsugu Nanba and Manabu Okumura. Producing more readable extracts by revising them. In Proc. of Coling, 2000.

[5] Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo and Genichiro Kikui. Optimizing Informativeness and Readability for Sentiment Summarization. In Proc. of ACL, 2010.

[6] Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo and Genichiro Kikui. Opinion Summarization with Integer Linear Programming Formulation for Sentence Extraction and Ordering. In Proc. of Coling, 2010.

[7] 奥村学, 難波英嗣. テキスト自動要約. オーム社, 2005.

[8] Chris D. Paice. Constructing Literature Abstracts by Computer: Techniques and Prospects. Information Processing & Management, 26(1):171–186, 1990.

[9] Ryohei Sasano, Daisuke Kawahara and Sadao Kurohashi. A Fully-Lexicalized Probabilistic Model for Japanese Zero Anaphora Resolution. In Proc. of Coling, 2008.

[10] 植田禎子, 荻野孝野, 飯田龍, 乾健太郎, 奥村学. 照応, 省略，共参照タグ付コーパスの構築. 言語処理学会第 11 回年次大会発表予稿集, 2005.