文節抽出型文簡約における読みやすさ向上のための文節末修正

福冨 諭　　高木 一幸　　尾関 和彦
電気通信大学
{fukutomi,takagi,ozeki}@ice.uec.ac.jp

1　はじめに

文章に対する自動要約として，文章中から重要と思われる文を抽出して並べることで要約とする文抽出型要約が広く知られている [1]．対象を文章から文に，抽出の単位を文から文節にすることで，文節抽出型文簡約の手法になる [2]．

文抽出では抽出の単位が比較的大きいため，文章中の重要な要素と冗長な要素とを十分に分離できない恐れがあるが，文簡約ではより小さい単位で抽出を行うため文章中の重要な要素と冗長な要素をより適切に分離し，重要な要素だけを抽出することが期待できる．

文抽出型要約の別の欠点として，原文では抽出した文の前にあった文が簡約文では省略されたときに，文の意味が正しく理解されなかったり読みにくかったりすることが指摘されており，要約文を修正することが提案されている [3]．

文節抽出型文簡約もこれと似た問題がある．係り受け関係にある 2 文節の係り文節だけが抽出され，受け文節が省略されたときに，もとは係り受け関係になかった文節との間に係り受け関係があるかのように見え，不自然な文となる．この問題に対する 1 つの解決法は簡約文中の文節ができるだけ自然な係り受け関係を持つように簡約文を生成することである [2]．しかし 1 つの文節を抽出するときに係り受け関係を成り立たせるような別の文節をも抽出しなければならない場合があり，文章中の重要な要素と冗長な要素を十分に分離できない恐れがある．

本研究では，原文中の係り受け関係になかったが，簡約の結果係り受け関係を持つようになった文節対に注目し，その係り文節を修正することで，簡約文中の受け文節と適切な係り受け関係を構築する手法を提案する．

修正には人手で作成したルールベースの修正規則と，コーパスを利用した統計ベースの修正規則とを組み合わせて用いる．

原文: 歯科衛生助手を / 務める / A さん
抽出した文節: 歯科衛生助手を / A さん
修正を加えた簡約文: 歯科衛生助手の / A さん

図 1: 文節の抽出と修正の例.

2　新しい係り受け文節対

原文では係り受け関係になかった 2 文節で，簡約文の文節として選ばれたことによって係り受け関係を持つようになったものを新しい係り受け文節対と呼び，その 2 文節間の関係を新しい係り受けと呼ぶ．このとき原文の係り受け構造が崩れたと表現する．原文の係り受け構造で新しい係り受け文節対に挟まれ，簡約文では省略された 1 つ以上の文節を特に省略文節と呼ぶ．

図 1 に例を示す．文節 “歯科衛生助手を” と “A さん” が新しい係り受け文節対であり，“務める” が省略文節である．

新しい係り受け文節対では受け文節を修飾できるように係り文節を修正する必要がある場合がある．図 1 の例では文節 “歯科衛生助手を” を “歯科衛生助手の” と修正している．原文では “歯科衛生助手を” は “務める” に係っていたが，簡約文では “A さん” に係るので修正する必要があった．

本研究では新しい係り受け文節対と省略文節の素性を用いて係り文節の主辞の活用形と付属語を修正する．

3　修正規則の生成

3.1　省略文節を利用したルールベースの修正規則の作成

文節 a が文節 b にかかり，文節 b が文節 c にかかるような 3 つの文節 a, b, c を考える．これらの文節のうち，a, c が簡約文に採用され，b が省略されたとする．

文節 a, c 間には新たな係り受けができる．このとき文節 a, b, c の情報を利用して a を修正する．

このような修正規則は人手で作成する．表 1 に例を示す．

3.2 統計ベースの修正規則のためのコーパスの統計処理

統計ベースの修正規則は自然性の推定と修正の 2 つからなる．新しい係り受け文節対の素性を持った文節対がコーパス中にある程度存在すれば，その文節対は自然だと推定する．修正することによって素性がコーパス中にある程度存在するのであれば修正を行う．

このような修正規則のためにコーパスの統計処理を行う．原文と簡約文を係り受け解析し，係り受け文節対の係り文節の主辞と主辞品詞，受け文節の主辞と主辞品詞の組み合わせについて，それぞれの出現回数を記録する．

4 修正処理

3 節の方法で生成した修正規則を利用して次のように修正処理を行う．

1. 省略文節の情報を利用し，新しい係り受けの文節対の係り文節を修正する．
2. 修正の有無に関わらず，新しい係り受け文節対の自然性を推定する．
3. 新しい係り受け文節対の自然性が低く，修正する候補があれば修正する．

この節の残りでは処理の各過程について述べる．

4.1 ルールベースの修正規則による修正

文末の文節から逆順にルールベースの修正規則を用いて修正を行う．適用できる修正規則がなければ修正は行なわない．

省略文節が複数ある場合の処理は次の 2 種類である．どちらの処理を適用するかは修正規則ごとに定める．

- 受け文節を修飾する省略文節の素性を利用する．
- 複数の修正規則を順に適用する．

原文: 歯科衛生助手を / 務める / A さんが / 行方不明
抽出した文節: 歯科衛生助手を / 行方不明
修正した簡約文: 歯科衛生助手が / 行方不明

図 2: 受け文節を修飾する省略文節を用いた修正の例.

原文: 弁護士， / A 容疑者が / 詐取した / 事件
抽出した文節: 弁護士， / 事件
途中経過: 弁護士， / A 容疑者の / 事件
修正した簡約文: 弁護士の / 事件

図 3: 複数の修正規則を順に適用した修正の例.

係り文節と受け文節を修飾する省略文節の素性から係り文節を修正できる場合には修正する．

図 2 に例を示す．この例では係り文節 “歯科衛生助手を” と受け文節を修飾する省略文節 “A さんが” はそれぞれの主辞品詞から “A さんという人物の職業・肩書が歯科衛生助手” という関係がわかるので，係り文節を “歯科衛生助手が” に修正する．

係り文節と受け文節を修飾する省略文節の素性のみからはその 2 文節の関係がわからない場合には，その間の省略文節に対して順に修正規則を適用する．

図 3 に例を示す．この例では係り文節 “弁護士，” と受け文節を修飾する省略文節 “詐取した” はそれらの文節のみからでは，“弁護士から搾取した” や “弁護士が搾取した” などがの関係が考えられ，一意に定まらない．そこでまず “A 容疑者が / 詐取した / 事件” という部分に注目し，“A 容疑者の / 事件” と修正する．次に “弁護士， / A 容疑者の / 事件” を “弁護士の / 事件” と修正することで最終的な簡約文を得る．

4.2 新しい係り受け文節対の自然性の推定

文節 a, c を新しい係り受け文節対とする．学習データに存在する (a の素性, c の素性) の出現回数が閾値よりも大きければ，a, c の係り受け関係は自然だと推定する．

素性は a, c のそれぞれについて主辞の文字列と主辞品詞を利用する．素性の組み合わせは 4 通りとなり，

表 1: ルールベースの修正規則の例.

係り文節 a の素性	省略文節 b の素性	修正後の a の付属語	例
サ変名詞 + “が”	サ変名詞	b の付属語	滞納が/発覚した/A 市長 → 滞納した/A 市長
名詞 + “が”	動詞 (連体形)	“の”	A 氏が/撮影した/写真 → A 氏の/写真
名詞 + “の”	固有名詞	b の付属語	歯科衛生助手の/A さんが → 歯科衛生助手が

それぞれについて閾値を設定する.

全ての組み合わせで出現回数が閾値を下回ったときは，係り受け関係が不自然だと推定する.

4.3 統計ベースの修正規則による修正

自然性の推定により，不自然と判定された文節対について，統計ベースの修正規則を用いて修正を行う.

文節 a, c を新しい係り受け文節対とし，文節 a' を a と同じ素性を持った学習データ中の文節とする．学習データに存在する (a の素性, c の素性, a' の文節末情報) のうち，出現回数の最も多いものを a の修正候補とする.

素性は a, c のそれぞれについて主辞の文字列と主辞品詞を利用する．素性の組み合わせは 4 通りとなり，それぞれについて閾値を設定する.

主辞の文字列を優先して比較を行い，出現回数が閾値よりも大きいときに a の文節末を a' の文節末に置き換えることで修正を行い，以降の比較は打ち切る.

5 読みやすさ向上のための補助的な処理

本研究の主題ではないが，読みやすさを向上させるために括弧の処理と文末の処理を行なった.

5.1 括弧の処理

括弧や鍵括弧の内容が複数の文節に渡り，開き括弧や閉じ括弧の文節が簡約文に採用されなかったとき，残った開き括弧や閉じ括弧が不自然さを感じさせる.

一組の括弧のうち開き括弧だけが簡約文に採用された場合には，括弧内の文節で簡約文に採用された最初のものに開き括弧を追加する．閉じ括弧についても同様の処理を行う.

5.2 文末の処理

文末の一部の助詞と動詞 “する” を削除し，活用形を基本形にする.

6 実験

修正によって簡約文の読みやすさが向上することを確認するために主観評価実験を行なった.

6.1 実験条件

毎日新聞全文記事および 54 文字データベース (2002 年度版)[4] の各記事から本文の第 1 文と記事の要約を抽出した．そのうち 200 記事から人手でルールベースの修正規則を作成し，29622 記事を統計ベースの修正規則のための学習データとした．333 記事で主観評価実験を行なった.

係り受け解析には茶筅 [5] と南瓜 [6] を用いた．主観評価実験のための記事では人手で係り受けの修正を行なった上で，本文の第 1 文と記事の要約とを比較し，第 1 文から要約記事に情報が利用されている文節を抽出した．これを出現順に並べ，5 節で述べた処理を適用したものを文節抽出による簡約文として用いた.

333 記事のそれぞれについて，文節を抽出して並べただけの簡約文，システムで修正した簡約文，人手で修正した簡約文の 3 文を作り，全ての簡約文をランダムに並び換えた．その各文節が不自然だと感じたときにそれを指摘し，自然になるように修正する課題を 10 名の被験者に与えた.

統計ベースの修正規則で，ある係り受けが自然であるかどうかの判定に用いる出現回数の閾値は，係り文節，受け文節ともに品詞を用いる場合には 3，それ以外の場合には 1 とした．不自然だと判定されたときの修正規則ための出現回数の閾値は，係り文節，受け文節ともに文字列を用いるときには 3，片方が品詞の場合には 500，双方が品詞の場合には 1000 とした.

表 2: ルールベースの修正規則の主観評価結果.

		自然な文節数	
省略文節	適用回数	修正前	修正後
連続する名詞	94	26	56
補足表現	89	8	58
固有名詞	36	8	21
動詞	33	3	12
並列表現	19	3	13
数量表現	9	2	5
形式名詞	5	0	5

表 3: 統計ベースの修正規則の主観評価結果.

素性			自然な文節数	
係り文節	受け文節	適用回数	修正前	修正後
文字列	文字列	43	12	21
文字列	品詞	0	0	0
品詞	文字列	1	0	1
品詞	品詞	8	1	2

6.2 結果

修正前の文について不自然であると指摘された文節数は被験者 10 名の平均で 223.1 文節，システムが修正した文については 138.3 文節，人手で修正した文については 45.3 文節であった.

係り受けが崩れていない文節は 1239 あり，そのうち 1143 文節が自然であると評価された．文末の文節は付加的な処理として読みやすくするための修正を行なったが，333 文のうち 279 文について自然であると評価された.

修正の前後で自然であると判断された文節の数を，最終的にルールベースの修正規則を適用したものについて表 2 に，統計ベースの修正規則を適用したものについて表 3 に示す.

統計ベースの修正規則のうち適切なものがなかったために修正を行わなかったものが 27 文節あり，そのうち 13 文節は自然であると評価された．統計ベースの修正規則と同様な，新しい係り受け関係を利用した，人手による修正規則を適用した文節は 29 文節あり，修正前には 23 文節，修正後には 25 文節が自然であると評価された.

6.3 考察

適用される文節がある全ての修正規則について，修正の結果，自然だと評価される文節数が増加している．全体としても不自然であると指摘された文節数が減少している，

係り受けを利用することで文節の自然さを向上させるような文節末の修正ができたことを示している.

7 まとめ

文節抽出型文簡約において，原文の係り受け関係を利用することで簡約文に新しく生じる係り受け関係の係り文節末を修正し，読みやすさを向上させた.

今後の課題には新しい係り受け以外の関係を利用した修正が挙げられる．例えば “A 氏が/B 大統領の/特使として/C 首相と/会談” から文節を抽出して “特使として/C 首相と/会談” とした場合には係り受け関係は崩れていないので新しい係り受け関係では扱うことができない．この文は “特使が/C 首相と/会談” と修正する必要がある．簡約文の主語や述語がどの文節なのかを同定することで，このような修正が行えるだろう.

謝辞

本研究の一部は文部科学省科学研究費補助金基礎研究 (C)16500077 の支援を受けた.

参考文献

[1] H. P. Luhn, “The automatic creation of literature abstracts,” IBM Journal of Research and Development, Vol.2, No.2, pp.159–165, 1958.

[2] R. Oguro, H. Sekiya, Y. Morooka, K. Takagi, K. Ozeki, “Evaluation of a Japanese sentence compression method based on phrase significance and interphrase dependency,” Proc. TSD 2002, pp. 27–32, 2002.

[3] H. Nanba and M. Okumura, “Producing more readable extracts by revising them,” Proc. COLING-2000, pp. 1071–1075, 2000.

[4] “毎日新聞全文記事および 54 文字データベース (2002 年度版),” 毎日新聞.

[5] 松本裕治他: “日本語形態素解析システム『茶筌』version 2.3.3 使用説明書,” 2003.

[6] 工藤拓, 松本裕治: “チャンキングの段階適用による日本語係り受け解析,” 情報処理学会論文誌, Vol.43, No.6, pp.4834–1842, 2002.