多重修飾に着目した文内要約：削除型換言

大竹 清敬*　　増山 繁
豊橋技術科学大学 知識情報工学系
kohtake@slt.atr.co.jp　　masuyama@tutkie.tut.ac.jp

1　はじめに

多重修飾された名詞句に着目した文内要約の手法について報告する．また，本研究で提案する手法を実装し，参加した NTCIR-2 のタスクのひとつである TSC の A-2 サブタスクでの結果の一部についても報告する．

計算機およびネットワークの発展によって，膨大な量の文書が分散されて蓄積されるようになった．我々は自らが抱える問題を解決するために，これらの文書から必要な情報を探しだし，利用しなければならない．一方で，有史以来生物としての人間の情報処理能力はほとんど変化していない．そのため，自動要約技術などにより，読み手が読む文書の量を制御できることが求められている．また，このような自動要約が可能になると，情報検索の検索効率を向上させることが可能となる．それは，検索過程において検索結果の文書を読むか否かの判断をするためにその文書の要約を用いることが可能になるためである．

単一の文書に対する要約研究は，長い歴史を持っており [1]，重要な文を選ぶ，重要文選択型の要約が多く行われてきた．しかしながら，文単位の要約では情報が欠落してしまう場合がある．また，文の重要度のみに基づく文選択の場合は文間の結束性が考慮されないため，非常に読みにくい要約になる可能性がある．そこで，情報の欠落を極力抑えるため，文を短かく表現しなおす要約手法が必要である．

本研究は，1 文単位の一般的な要約の生成を目的とする．「質の良い文章要約を行うためには，ある一つの言語現象だけをとらえた談話解析だけでは不十分である」[2] という山本らの主張が示すように，質の良い要約を行うためには様々な言語現象に対応しなければならない．これは文章全体の要約はもとより，1 文での要約においても言える．しかしながら，文内での要約を考えた場合に，構文的そして意味的に重要度が低いと考えられる部分を削除することは自然である．そこで，本研究では，名詞にかかる修飾要素を削除することによって要約を作成する．

文内での要約を考えた場合に，不要箇所を削除して要約を作成する他に，換言によってより短い表現とする方法が考えられる．このような研究として文献 [3, 4] が挙げられる．これらのアプローチの問題点は，規則を人手で用意しなければならないことと，要約率[1]にそれほど貢献しない点である．また，このような換言規則を原文と要約結果とを比較することにより自動的獲得する手法 [5] も提案されているが，知識獲得に必要な要約済み文書が大量に存在しないのが現状である．

文内の不要箇所を削除して要約を作成する場合に，不要箇所の削除によって構文構造が破壊されると文の意味が伝わらなくなるため，構文解析が必要となる．日本語を対象とし，なおかつ構文解析結果を積極的に利用した文内要約に関するこれまでの研究は少ない．その理由は，構文解析器の精度が低く，また，大量の文書を実用的な時間で解析できるほどに計算機能力が高くなかったためと考える．近年の計算機および自然言語処理技術の向上により，大量文書の構文解析を十分実用的な速度で，そしてかなり高い精度で行えるようになりつつある．

三上ら [6] はニュース原稿を対象として構文解析器を用いて文内要約を行っている．しかし，そこでの構文解析器の利用は三上ら自らが簡易構文解析と述べているように積極的に構文解析結果を利用しているとは言い難い．石﨑ら [7] も構文解析器を利用しているが，石﨑らが構文解析器を用いている理由は，係り受け関係を調べるためであり，構文構造を直接利用して要約を生成しようとする本研究とは利用目的が異なる．また，直接要約を生成するわけではないが，亀田 [8] は簡易日本語解析系 QJP を用いて文の骨格となる文節群を強調表示する機能を提供している．構文的に重要度が低いと考えられる部分があり，それを利用する基本的な考え方は本研究と同一である．

本研究では，構文解析結果を積極的に利用し，要約を

*現在，ATR 音声言語通信研究所

[1]本研究における要約率は $\frac{\text{要約後の文章の長さ}}{\text{元文章の長さ}}$ である．

生成することを試みる．構文解析結果を用いて不要箇所を削除し，要約を作成しようという場合，直観的に不要と考えられるのは連体修飾成分である．山本らのシステム GREEN[2] や，三上ら [6] の要約手法においても連体修飾成分を削除することにより文内要約を実現している．しかしながら，三上ら [6] は，連体修飾節の削除，固有名詞へ係る修飾語句の削除のいずれにおいても，重要部の認定が困難であると報告している．

連体修飾節の要約における残存傾向は，文献 [9] に詳しい．ザトラウスキーは，寺村 [10] が言う「内の関係」ならびに「外の関係」という分類を持ち込んだ場合に，内の関係は要約に残存しにくく，外の関係 (特に内容的な) は残存しやすいという報告をしている．しかしながら，ある連体修飾節を要約に含めるべきか否かの判断は，文脈や，修飾成分と被修飾語との関係を無視してできるものではなく,「連体修飾節が内の関係ならば削除する」などのルールだけでは質の良い要約はできない．

連体修飾成分の削除による文内での要約を考える場合に，ある名詞を修飾する単体の連体修飾成分を削除するアプローチは不自然な要約を作成する可能性があると考える．ただし，これは被修飾名詞に依存している．たとえば，固有名詞のようにかなり限定された具体的なものを指す名詞は連体修飾成分を削除しても不自然に感じない場合がある．反対に，内容節などによりその被修飾名詞の内容を説明するような修飾成分を削除すると元の被修飾名詞が何を指しているかがわからず，不自然である．このような被修飾名詞とその修飾成分の関係は，被修飾名詞のどのような側面を修飾成分が修飾しているかを考慮しないかぎり適切な削除は行えない．また，適切な削除という点では，被修飾名詞がさらに，その文のなかでどのような役割を担っているかも考慮しなければならない．たとえば，さらに「名詞の〜」という形で他の名詞を修飾しているのか，あるいは格助詞を伴ないある動詞の格要素となっているのか，などである．そのような被修飾名詞と修飾成分の関係を，今後詳細に調査する必要はあるが，本研究では，ある名詞に対する修飾成分が複数存在するような場合には，修飾成分を削除しても比較的不自然にならないという現象に着目し，修飾成分の削除による要約を試みる．

2　二重修飾に着目した文内要約

2.1　多重修飾

文にみられる修飾関係には，さまざまな関係があるが，本研究では名詞を被修飾要素とする場合を中心的に扱う．ある名詞を修飾する連体修飾成分には，表 1 に示す種類がある [10].

表 1: 連体修飾の種類

文法的性質	例
こそあど詞連体形	この話
連体詞	ある話
形容詞の現在形・過去形	おもしろい話
形容動詞の連体形・過去形	変な話
名詞＋接続助詞「の」	昔の話
名詞＋格助詞＋「の」	昔からの話
名詞＋取立て助詞＋「の」	ここだけの話
副詞＋「の」	突然の話
関係節	私が聞いた話
内容節	子狸が少年と仲良くなる話

関係節：被修飾名詞が修飾節の格要素 (内の関係)
内容節:被修飾名詞が修飾節の格要素ではない (外の関係)

これらの連体修飾成分が，一つの名詞に対して複数修飾している状態を山本ら [2] の定義に従い多重修飾とする．特に，修飾成分が 2 つの場合を二重修飾と呼ぶ.

「X は」の形で文の陳述の対象を表す要素を,「主題」と呼ぶ [11]．文内の要約を行う上で，この主題が関係してくるため，ここで説明しておく．主題を提示する働きをする助詞を提題助詞と言い，一般に主題は名詞と提題助詞で構成される．提題助詞には,「は，なら，って，ったら」がある [11]．それに対して，同類の他の事項を背景にして，ある事項を取り上げる働きをする助詞を,「取り立て助詞」と呼ぶ．取り立て助詞には,「は，も，さえ，でも，すら，だって，まで，だけ，ばかり，のみ，しか，こそ，など，なんか，なんて，くらい」がある．

本研究では主題部を定義し，この主題部内から主題部の外へ係る依存関係は無視する．その理由は，次の 2 つによる．ひとつは，そのような依存関係は，その距離が長くなる傾向があり，構文解析器が誤る可能性が高くなるためである．もうひとつは，主題部内のある文節が主題部の外の名詞に対する二重修飾の修飾要素として認定され，削除された場合，主題の一部が欠落する．そのような場合，文の大意が損なわれる可能性があるからである．ただし，主題部の中で完結している二重修飾については，その他の通常の二重修飾と同様に扱う．

本研究において主題を表わすための提題助詞は，取り立て助詞を指すものとする．主題部を以下に定義する．

定義 文頭から，取り立て助詞を末尾にもつ文節までを「主題部」とする．また，取り立て助詞のあとに格助詞が続く場合もこれに該当する．

文のなかには，主題部を持たないものも当然あり，これを無題文という．

2.2 要約手法

本研究では，連体修飾成分を削除することによる要約を考える．要約結果を可能な限り自然に保つために，多重修飾に着目する．しかしながら，一般に修飾成分を3つ以上持つ名詞が使用されるような事は極めて稀であり，そのような場合は連用形となる場合が多い．たとえば,「白い長い大きな手」よりは,「白くて長い大きな手」が用いられる．以上の事実と，二重修飾に対する処理を応用することで多重修飾の場合も文内要約処理が可能であると考え，本研究では二重修飾のみを扱う．

要約を行うために連体修飾成分を削除することは自然であるように，そして構文解析器を利用すると容易に実現できるように考えられるが，ニュース文を対象とした実験でその難しさが確認されている [6]．連体修飾成分を削除した結果が不自然に感じられる場合は，削除した連体修飾成分が名詞を限定する働きが強い，あるいはその名詞の内容を示していると考える．三上らは，修飾成分に含まれる重要語や，被修飾要素の名詞によって不自然な削除を避けようとしたが，結果は不十分であったと報告している．

一方で，ある名詞に対してその修飾成分が二つ存在する場合，どちらか一方を削除しても，その名詞が裸になるわけではないため文の大意が大きく変化することはないと考える．そこで，我々は，二重修飾の場合にどちらか一方の修飾成分を削除し，要約を作成する．山本らの GREEN[2] でも，同様に多重修飾の場合に修飾成分を削除し，要約を作成している．山本らの手法では，最終の修飾成分を残し，他を削除する．しかしながら，この方法では,「ここに規制緩和の大きな役割がある.」を要約した場合に,「ここに大きな役割がある.」となり不自然である．この場合は,「ここに規制緩和の役割がある.」とする方が自然だと考える．

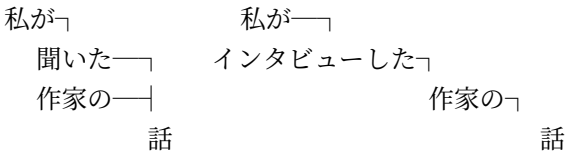
日本語では，長くて複雑な構造をもった成分を，文の前方に置こうとする傾向がある [12]．さらに，長くて複雑な構造をもつ連体修飾成分は，名詞を限定する働きが強いと考える．したがって，本研究では，名詞を限定する働きの強さを基準としてどちらの修飾成分を削除するかを決定する．名詞を限定する働きが同程度の場合は，GREEN と同様に前方の修飾成分を削除する．ただし，削除した結果が不自然になると考えられる場合は，削除を行わない．この削除の回避は，二重修飾の修飾成分と被修飾要素との間に制約を設けることにより実現する．二重修飾の前方と後方の修飾成分ならびに被修飾要素の名詞を含む文節が与えられたときの削除規則の例を表 2 に示す．なお表中の “—” は，特に制約がないことを意味する．このような規則が合計 36 個存在する．

表 2: 二重修飾の削除規則例

前方	後方	被修飾要素	動作
～の	～の	—	削除しない
—	—	～との	削除しない
～という	～の	～の	後方を削除
連体修飾節	形容詞	—	後方を削除
—	～な	こと…	削除しない

2.3 二重修飾の特例

二重修飾を扱う場合の特別な例として,「連体修飾節+名詞の+名詞」という構造を考える．たとえば「私が聞いた作家の話」と「私がインタビューした作家の話」を考える．両者は，各形態素の品詞は全く同じ列を構成するが， その依存構造は下に示すように異なる．



このような，依存構造の異なる文に対しては，構文解析器が解析を誤る傾向が強くなる．このような場合の対処として山本らの GREEN では，いずれの場合も「名詞の」の部分が被修飾要素である名詞にかかることから，こちらを残し，連体修飾節を削除している．しかしながら，連体修飾節が「名詞の」の内容を説明しているような場合は，連体修飾節を削除すると文の大意を損う可能性が大きくなる．そこで,「連体修飾節+名詞の+名詞」は一般の 2 重修飾の処理とは別に処理を行う．基本的には，山本らの GREEN と同様に連体修飾節を削除する．ただし，文の大意を保つため「名詞の+名詞」という構造のみでは被修飾名詞の内容が十分に伝わらない場合は連体修飾節を削除しない.「名詞の+名詞」という構造のみで被修飾名詞の内容が十分に伝わるかどうかは「名

詞の」に含まれる名詞が抽象的すぎないかどうかで判断する．これは，三上らが形式的表現と呼んだものに相当する．抽象的かどうかを判断するために，日本語語彙大系 [13] の各カテゴリを代表する名詞を抽象的な名詞として利用した．

3 評価実験

本研究で提案する文内要約手法を実装し，評価実験を行った．使用した形態素解析器は JUMAN，構文解析器は KNP である．実装は *Perl* を用いて行った．

評価実験は，NTCIR-2 のタスクである TSC の A-2 サブタスクに参加することで行った．A-2 サブタスクは，人間が作成した要約と比較可能な要約を作成するものである．A-2 サブタスク参加にあたり，今回報告する二重修飾に着目した要約手法だけでは，冗長さが残存した要約となるため，以下の文内要約手法もあわせて実装した．以下，簡単に説明する．

補足説明の削除 括弧などで表現される付加的な情報を削除する．

例：公職選挙法（一九九条）違反 → 公職選挙法違反

直接引用表現内の文の削除 直接引用表現が複数の文で構成されるとき，最初の文を削除する．ただし，2 文目以降に一文を参照する可能性のある指示詞が存在する場合は，削除を行わない．このような削除を行った場合，記事を理解する上でも，削除後の自然さの点でも大きな問題となることはないことを経験的に確認している．

例：そのうえ「明日の公式試合には出なくてええ．背番号も返せ」と言われたという．→そのうえ「背番号も返せ」と言われたという．

直接引用表現の削除 児玉ら [14] が指摘したように，ある直接引用表現の後には，その要約が続く場合がある．これに該当する場合は直接引用表現を削除する．

例：検察側は「捜査段階で事実を認めていた」と主張して，タイミングを計って証拠申請する構え．→ 検察側はタイミングを計って証拠申請する構え．

例示の削除 「〜などの+名詞」という表現において「〜などの」を削除する．ただし,「〜などの」が係る名詞が抽象的すぎる場合は削除しない．また,「〜などで」が用言に単独で係る場合,「〜などで」を削除する．

例 1：経済や外交戦略などの専門知識はもとより，→専門知識はもとより，

例 2：既に蔵相・外相会合などで取り上げられている→既に取り上げられている

テーブルによる換言処理 文頭の接続詞や，文末表現などを換言テーブルを用いて削除する，あるいは，短い表現に変換する．新聞記事用に規則を人手で 106 用意した．

例 1：…決まらないようだ． → …決まらない．

例 2：そんな中，…． → …．

3.1 要約システム

上述した文内要約手法だけでは，A-2 サブタスクにて規定される文字数まで要約することは困難なことから，文選択による要約手法を用いた [15]．したがって，実験で用いた要約システムは，文選択による要約手法の結果に対して本研究で提案する文内要約の手法を適用することにより要約を行う．要約システムの処理の概要を以下に示す．

1. 文選択による要約手法により各文の重要度を決定する．
2. 各文に対してテーブルによる換言処理を除く文内要約を行う．
3. 規定された文字数を越えるまで，重要度が大きい順に文を選ぶ．
4. 全ての文に対してもう一度，テーブルによる換言処理を除く文内要約を行う．
5. 選択した文全ての長さが，規定された文字数以下ならば要約結果を出力して終了．
6. 選択した文に対してテーブルによる換言処理を行う．
7. 選択した文全ての長さが，規定された文字数以下ならば要約結果を出力して終了．
8. 選択した文の中で最も重要度が低い文を除き，規定文字数 – 現在の文字数以下の長さの文をまだ選択されていない文のなかから重要度が高い順に探す．
9. 該当する文があれば，それを採用し，テーブルによる換言処理を行い，要約結果を出力して終了．
10. 該当する文が存在しない場合，8. の処理を終えた段階で選択されている文を要約として出力する．

3.2 要約結果

実験の対象とした文書は毎日新聞 1994 年から 15 記事，1998 年から 15 記事選択された A-2 サブタスクの課

題記事合計 30 記事である．30 記事全ての文に対して文内要約を行った場合の平均要約率を表 3 に示す．これは，換言処理を除く文内要約を 2 回行った場合の結果と，それにさらにテーブルによる換言処理を行った場合の結果である．なお，ここに示した結果は，記号だけからなる文などをそのまま含めて要約を行ったものである．

表 3: 全 30 記事の要約結果

	換言処理なし 2 回	換言処理追加
要約率	93.01%	91.68%

4 評価

本研究で提案した手法ならびに，それを用いた要約システムを評価するために，NTCIR-2 のタスクである TSC の A-2 サブタスクに参加し，要約システムを評価した．A-2 サブタスクでは 2 種類の評価を行った．一つは主観評価である．主観評価では，記事ごとに「文書として読みやすいか」と「元の文書の重要な内容を不足なく記述しているか」の 2 つの観点から各要約を評価する．もう一方は，content-based での評価である．content-based な評価では，人間の作成した要約およびシステムの作成した要約をともに，形態素解析し，内容語のみを抽出する．そして，人間の作成した正解要約の単語頻度ベクトルとシステムの要約の単語頻度ベクトルの間のコサイン距離を計算し，どの程度内容が単語ベースで類似しているかという値を求める．

A-2 サブタスクで評価を行ったのは，指定された全 30 記事をそれぞれ 20%と 40%に要約したものである．

主観評価では，読みやすさと内容の 2 点についてそれぞれ良いものから 1,2,3,4 の 4 段階で評価値を付ける．つまり，評価値が小さければ小さいほど良い要約となる．評価結果を表 4 に示す．なお，評価値の平均の括弧内の数値は参加した全 10 システムの平均を表す．

表 4: 主観評価結果

要約率 (%)	観点	評価値の平均
20	読みやすさ	2.53(3.16)
20	内容	2.93(3.24)
40	読みやすさ	2.73(3.05)
40	内容	2.77(3.12)

content-based な評価では，比較対象とする要約として人間が自由に作成した要約と，人間が重要箇所を抽出して作成した要約の 2 種類がある．ベクトルの要素は，各内容語（名詞，動詞，形容詞，未定義語）の $tf \cdot idf$ 値である．また，ベクトル間の類似尺度として，コサイン距離を用いる．評価結果を表 5 に示す．なお，評価値の括弧内の数値は参加した全 11 システムの平均を表す．

表 5: content-based な評価結果

要約率 (%)	比較対象	評価値
20	自由作成	0.4727(0.4418)
40	自由作成	0.6483(0.6065)
20	重要箇所抽出	0.5137(0.4740)
40	重要箇所抽出	0.6608(0.6342)

また，A-2 サブタスクの課題記事を 20%と 40%に要約した場合に文内記事要約のために要約システムが用いた各要約手法の総適用回数の内訳を表 6 に示す．つまり，表 6 は，課題記事の 20%と 40%の要約を作成するために用いた各要約手法ののべ回数の内訳を示している．

5 考察

まず，表 3 に示した実験結果から本研究で提案した二重修飾に着目した要約手法は，要約率に大きな貢献していないことがわかる．その理由は，不自然になる要約を極力回避したためであると考える．一方，換言テーブルを用いた換言処理による要約は，文の大意を損ねることなく，安全に要約を行えると言える．しかし，テーブルの作成にコストがかかる点が問題である．

本研究で提案した文内要約手法を，文選択型の要約システムの結果に適用し，TSC の A-2 サブタスクにて評価した結果は，相対的に良好な結果を示した．content-based な評価では，本研究で提案する手法のみでは，要約率を小さくできないところから，文選択型の要約システムの性能に依存することは否定できない．しかしながら，主観評価における読みやすさの評価が相対的に良かったことは，要約結果を極力自然に保とうとした点から評価できる．

表 6 から，二重修飾による削除が削除文字数の構成比上最も大きく，要約率を小さくすることに貢献していることがわかる．一方で，実際の要約結果においては，

表 6: 各要約手法の内訳

手法	適用回数	総削除文字数	文字数の構成比 (%)
補足説明の削除	335	662	31.1
直接引用表現内の文の削除	7	277	13.0
直接引用表現の削除	1	21	1.0
二重修飾の一方の成分の削除	61	729	34.2
例示の削除	6	126	5.9
換言処理	89	314	14.7

二重修飾のうち,「連体修飾節+名詞の+名詞」という構成の連体修飾節を削除した場合に不自然に感じられる場合が多いという印象を持った．この点に関して，厳密に評価し，改良することは今後の課題である．

6 むすび

本研究では，二重修飾に着目した要約手法を提案した．提案した要約手法を含む要約システムを実装し，NTCIR-2 のタスクである TSC の A-2 サブタスクによって要約システムを評価した．その結果，良好な結果を示した．一方で，極力自然さを保つことを目標として，二重修飾に着目した要約手法をとり込んだ要約手法のみでは，要約率が小さい要約を生成することは困難であることがわかった．

参考文献

[1] Mani, I. and Maybury, M. T. eds.: *Advances in Automatic Text Summarization*, MIT press (1999).

[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] 若尾孝博, 江原暉将, 白井克彦：テレビニュース番組の字幕に見られる要約の手法, 情報処理学会研究報告 97-NL-122, pp. 83–89 (1997).

[4] 山崎邦子, 三上真, 増山繁, 中川聖一：聴覚障害者用字幕生成のための言い換えによるニュース文要約, 言語処理学会 第 4 回 年次大会発表論文集, pp. 646–649 (1998).

[5] 加藤直人:ニュース文要約のための局所的要約知識獲得とその評価, 電子情報通信学会信学技報 NLC98-7, pp. 7–14 (1998).

[6] 三上真, 増山繁, 中川聖一：ニュース番組における字幕生成のための文内短縮による要約, 自然言語処理, Vol. 6, No. 6, pp. 65–81 (1999).

[7] 石塚友子, 片岡明, 増山繁, 山本和英, 中川聖一：係り受け関係を用いた重複表現削除, 自然言語処理, Vol. 7, No. 4, pp. 119–142 (2000).

[8] 亀田雅之：日本語文書読解支援系 QJR の検討, 情報処理学会研究報告 NL-110, pp. 57–64 (1995).

[9] ポリー・ザトラウスキー：要約文における連体修飾の残存傾向, 佐久間まゆみ（編）, 文章構造と要約文の諸相, pp. 61–78, くろしお出版 (1989).

[10] 寺村秀夫:日本語の文法（下）, 大蔵省印刷局 (1981).

[11] 益岡隆志, 田窪行則：基礎日本語文法 –改訂版–, くろしお出版 (1992).

[12] 野田尚史：語順を決める要素, 言語, Vol. 29, No. 9, pp. 22–27 (2000).

[13] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦（編）：日本語語彙大系 CD-ROM 版, 岩波書店 (1999).

[14] 児玉充, 片岡明, 増山繁, 山本和英：直接引用表現を利用した要約知識の自動抽出の試み, 言語処理学会第 6 回年次大会 発表論文集, pp. 241–244 (2000).

[15] Ohtake, K., Okamoto, D., Kodama, M. and Masuyama, S.: Yet Another Summarization System with Two Modules Using Empirical Knowledge, in *Proceedings of NTCIR Workshop 2 Meeting* (2001), http://research.nii.ac.jp/~ntcadm/publication1-ja.html.