連体修飾表現の省略可能性に関するコーパスからの知識獲得

酒井 浩之　　増山 繁
sakai@smlab.tutkie.tut.ac.jp, masuyama@tutkie.tut.ac.jp
豊橋技術科学大学　知識情報工学系

1　はじめに

近年，テキスト自動要約の必要性が高まってきており，自動要約に関する研究が盛んに行なわれてきている [1]．要約とは，人間がテキストの内容の理解，取捨選択をより容易にできるようにするために，元のテキストを短く表し直したものをいう．

これまでの研究で提案されてきた要約手法は，重要な文を選ぶ重要文抽出型の要約や，一文ごとに要約を行なう文内要約などがある．しかし，どのような使用目的の要約でも作成できる万能な要約手法は存在しないため，要約の使用目的に応じた手法を選択し，時には複数の手法を併用して要約を作成することが必要となる [2]．

本論文では，不要箇所省略による要約を実現するための一手法である，文中の省略可能な連体修飾表現を認定するために必要な知識をコーパスから獲得する手法を提案する．不要箇所省略による要約手法として，大竹ら [3] は，一文ごとの要約ヒューリスティックスに基づいた二重修飾表現などの削除を提案している．若尾ら [4] や山崎ら [5] は，人手で作成された字幕とその元となったニュース原稿とを人手で比較し，それによって作成した言い換え規則を用いた要約手法を提案している．また，加藤ら [6] は記事ごとに対応のとれたニュース原稿と字幕放送の原稿を用いて，言い換えに関する要約知識を自動獲得する研究を行なっている．

ところが，これらの手法には次のような問題点がある．まず，不要箇所の省略や言い換えに関する規則を人手で作成するには多大な労力を要し，更に，網羅性などの問題も残ることが挙げられる．また，加藤らが使用したような原文と要約文との対応がとれたコーパスは要約のための言語知識を得る知識源として有用であるのは明らかであるが，一般には存在しておらず，入手するのが困難である．また，そのようなコーパスを人手で作成するには多大な作業量が必要であると予想される．

このような理由から，本論文では，原文と要約文との対応がとれていない一般のコーパスから，不要箇所省略による要約において利用できる言語知識を自動的に獲得し，獲得した言語知識を用いて要約を行なう手法を提案する．ここで不要箇所の単位として連体修飾表現の動詞連体修飾表現（以降，動詞連体とする）に着目する．連体修飾表現の省略は，例えば，大竹ら [3] が行なったが，大竹らの着目した連体修飾表現は，二重修飾における連体修飾表現であった．これは連体修飾表現の意味情報に立ち入らず，二重修飾という構文構造に基づいた要約規則を人手で作成して省略部分を認定しており，静的な省略箇所認定となる．そのため，人手で抽出した規則があてはまる用例にしか適用できず，網羅性の問題がある．それに対して，本論文ではコーパスの統計的情報や文脈情報を使用して，名詞の修飾のされやすさ，修飾の多様性，重要度，および，認定対象の動詞連体の重要度や冗長度といった指標を考慮にいれた，動的な省略可能な動詞連体の認定を行なう手法を提案する．本手法では，新聞記事などのコーパスから教師なしで要約知識を獲得できるので，従来手法における網羅性，コスト，要約文の入手困難といった問題が解決できる．

具体的には，省略できる可能性のある動詞連体が修飾している名詞に対して，“修飾されやすさ”，“修飾多様性”をコーパスから調べ，修飾される頻度が低い，もしくは，修飾する動詞の種類が限定されている名詞に係る動詞連体を省略可能と認定する．同時に，その名詞の重要度を算出し，重要な名詞に係る動詞連体を省略可能と認定する．また，その動詞連体の内容および前後の文脈を考慮して，その動詞連体に含まれている情報が以前の文にも含まれている情報である場合には，省略可能と認定されやすくなる．逆に，重要な情報が含まれている場合には省略可能と認定されにくくなるような工夫を行なっている．

本研究でコーパスとして想定するのは，形態素情報などの付与されていない一般のコーパスである．したがってCD-ROMなどで提供されている新聞記事のバックナンバーや電子辞書，WWW上で公開されている文書などを利用することができ，コーパスの大規模化も比較的容易に実現可能である．

以下，第２章では，本論文で提案する手法を説明する．第３章では，手法を実装して，それによって省略可能と認定される連体修飾表現を示す．第４章では，実験を行ない，本手法を評価する．第５章では，実験結果の考察を行なう．

2　提案手法

本手法では，省略できる可能性のある動詞連体が修飾している名詞に対して，“修飾されやすさ”，“修飾

多様性" をコーパスから調べ，修飾される頻度が低い，もしくは，修飾する動詞の種類が限定されている名詞に係る動詞連体を省略可能と認定する．修飾される頻度が低い名詞は修飾されなくても意味が分かる名詞であり，一方，たとえ修飾される頻度が高くても，修飾する動詞の種類が限定されている名詞に対する動詞連体は一般知識から内容が補完できる．よって，そのような名詞を修飾している動詞連体を省略可能としても情報欠落が少ないからである．同時に，その名詞の重要度を算出し，重要な名詞に係る動詞連体を省略可能と認定する．ここで，動詞連体とは，名詞を修飾する連体修飾表現の最後に動詞が出現し，その動詞が修飾する名詞に係っている連体修飾表現と定義する．例えば，「大舞台で演技するスターの気分」という文で，連体修飾表現「大舞台で演技する」の動詞「演技する」が「スター」に係っているので，「大舞台で演技する」は動詞連体である．以降，名詞 n を修飾する動詞 v をもつ動詞連体を $VP(v,n)$ と定義する．具体的には，動詞連体 $VP(v,n)$ に対して，以下の計算式で重みを算出し，重みが小さい動詞連体を省略可能と認定する．なお，名詞 n が複合名詞である場合，最後に出現する名詞を $l(n)$ とする．例えば，n=「市場統合」なら，$l(n)$ = 「統合」である．一般名詞なら $l(n) = n$ である．

$$W(VP(v,n)) = \frac{E(l(n)) \times M(VP(v,n))}{idf(n) \times J(n)} \times CR(VP(v,n)) \quad (1)$$

但し

$E(l(n))$: 名詞 $l(n)$ に係る動詞の確率に基づくエントロピー．詳細は後述するが，多様な動詞によって修飾される名詞であるほど高い値を算出するため，$W(VP(n))$ が高くなり，そのような名詞に係っている動詞連体は省略されにくくなる．

$idf(n)$: 名詞 n を idf[7] に基づいて計算した値．具体的には，以下のような数式になる．

$$idf(n) = \log \frac{N}{df(n)} \quad (2)$$

但し，

N: 対象コーパスにおける全文書の総数，

$df(n)$: 対象コーパスにおいて，名詞 n を含む文書の数，

$J(n)$: 名詞 n が複合名詞であった場合，それを構成している名詞の数，例えば，「市場統合」なら「市場」と「統合」が連結した名詞なので，$J(n)$=2 である．なお，複合名詞でない名詞の場合は $J(n)$=1 である．そのため，複合名詞の方が，$W(VP(v,n))$ の値が低くなり，係っている動詞連体が省略されやすくなる．

$M(VP(v,n))$: 動詞連体 $VP(v,n)$ の動詞 v に係っている連用修飾表現の数，例えば，動詞連体「これまでＥＣ市場を分断してきた」は，動詞「分断する」に「これまで」「ＥＣ市場を」の 2 つの連用修飾表現が係っているので，$M(VP(v,n)) = 2$ である．

$CR(VP(v,n))$: 動詞連体 $VP(v,n)$ において，内容および文脈を考慮した文脈補正項，詳細は後述する，

2.1 名詞 $l(n)$ に係る動詞の確率に基づくエントロピー

対象としている動詞連体が修飾している名詞に対して，"修飾されやすさ"，"修飾多様性" をコーパスから調べる．そのため，その名詞に対してコーパス全体で動詞ごとに修飾される確率を調べ，そのエントロピーを算出する．具体的には以下のようにエントロピー $E(l(n))$ を算出する．

$$E(l(n)) = -\sum_{v \in V(l(n))} P(v,l(n)) \log(P(v,l(n))) \quad (3)$$

但し

$V(l(n))$: 名詞 $l(n)$ を修飾する動詞連体に含まれる動詞の集合，

$P(v,l(n))$: 名詞 $l(n)$ が，動詞 v を含む動詞連体で修飾される確率，すなわち，

$$\frac{\text{名詞 } l(n) \text{ が動詞 } v \text{ を含む動詞連体で修飾される総数}}{\text{名詞 } l(n) \text{ が動詞連体で修飾される総数}}$$

すなわち，多様な動詞によって修飾される名詞（例えば「案」）は，確率 $P(v,l(n))$ があまり変化しないので，エントロピーが高くなる．

2.2 文脈補正項について

式 (1) の $CR(VP(v,n))$ は対象としている動詞連体において，内容および文脈を考慮した文脈補正項である．具体的には以下のような式になる．

$$CR(VP(v,n)) = 1 + \sum_{p \in P(VP(v,n))} B(p,VP(v,n))$$

$$B(p,VP(v,n)) = \frac{1 + after(p,VP(v,n))}{2(1 + before(p,VP(v,n)))} \times \log \frac{N}{df(p)} \quad (4)$$

但し

$P(VP(v,n))$: 動詞連体 $VP(v,n)$ に含まれる名詞の集合，ただし，複合名詞の場合は分解せずに複合名詞を 1 つの名詞として扱う．

$after(p,VP(v,n))$: 名詞 p を含む動詞連体 $VP(v,n)$ より後の文に名詞 p が出現する頻度．ただし $p \in P(VP(v,n))$，

$before(p,VP(v,n))$: 名詞 p を含む動詞連体 $VP(v,n)$ より前の文に名詞 p が出現する頻度，

省略可能であるかどうかの認定対象となっている動詞連体に，それより以前の文に出現した名詞が含まれている場合は値が小さくなる．そのため，そのような動詞連体は省略可能と認定されやすくなる．また，重要な名詞を多く含む，長い動詞連体は多くの情報を含み，それを省略することで情報欠落が大きくなる危険がある．しかし，重要な名詞を多く含む動詞連体の値は，文脈補正項によって高くなるので省略可能と認定されにくくなる．

表 1: 制約として設定した名詞

こと	もの	わけ	上	中	他	ほか	前	後
間	際	うえ	ため	くらい	ところ	よう		
かぎり	必要	動き	一方	下	向き	なか		

2.3 その他の制約

本手法では，精度を上げるために以下の 3 つの制約を設ける．

制約 1 $CR(VP(v,n)) = 1$ である動詞連体 $VP(v,n)$ は省略不可とする．

制約 2 動詞連体 $VP(v_1,n_1)$ の最初に出現する名詞に別の動詞連体 $VP(v_2,n_2)$ が係っている場合，動詞連体 $VP(v_1,n_1)$ を省略不可とする．

制約 3 名詞 n に対して制約を設けて，制約として設定した名詞が修飾されていた場合，その動詞連体は省略不可とする．

$CR(VP(v,n)) = 1$ とは，動詞連体に名詞が含まれていない場合に起こる．そのような動詞連体を省略すると情報欠落は大きくないが，文間のつながりが壊れる場合が多く，制約 1 によって省略不可とする．

制約 2 は，もし，$VP(v_1,n_1)$ を省略可能とすると，$VP(v_2,n_2)$ も省略可能と認定する必要がある．そのため，省略箇所が多くなり情報欠落が大きくなるためである．

制約 3 は，例えば，一文字で構成される名詞など，それだけでは意味を成さない名詞を人手で判別し設定する．表 1 に制約として設定した名詞を示す．

3 手法の実装

本手法を実装して，文書の要約システムを作成した．コーパスは 93 年の日経新聞記事 1 月 1 日から 6 月 30 日までの，66686 記事を採用した．形態素解析器として JUMAN version 3.5 を，構文解析器として KNP version2.0b6 を採用した．実際には，本手法によって省略可能と認定された動詞連体を文から削除することによって，削除型の文内要約を実現することができる．

4 評価実験

実装したシステムを評価した．実験における対象記事は，日経新聞 93 年の 1 月 1 日から 6 月 30 日までの 66686 記事の中から無作為に 20 記事を選択した．選択した 20 記事には全部で 174 の動詞連体が存在した．これが認定対象となる．

本手法によって 174 の動詞連体から省略可能な動詞連体を認定した．評価方法は，対象記事群から省略可能な動詞連体の正解データを作成し，適合率，再現率で性能を評価する．ここで正解データは，対象記事群における全ての文から，省略しても妥当な動詞連体を人手で抽出し，作成した．再現率，適合率の定義を示す．

表 2: 性能評価

手法	A	B	C
認定数	67.0	134	81
再現率 (%)	72.7	90.4	72.6
適合率 (%)	79.3	49.3	65.4

$$\text{再現率} = \frac{\text{本手法による結果と正解データで一致する数}}{\text{正解データの省略可能な動詞連体の数}}$$

$$\text{適合率} = \frac{\text{本手法による結果と正解データで一致する数}}{\text{本手法によって省略可能と判定された動詞連体の数}}$$

実験は，以下の 3 つの手法に対して，それぞれ実験結果をとり，再現率，適合率を算出して評価する．

手法 A: 本手法，

手法 B: 2.3 節で述べた制約 1〜制約 3 に該当しない動詞連体を全て省略可能とする手法，

手法 C: 式 (1) の文脈補正項を導入しない手法，

手法 A は閾値によって結果が変動する．そのため，再現率と適合率による F 値が最も大きい閾値を中心に閾値を変化させ，個々の再現率，適合率の平均を採用する．その結果，閾値 4.1 で F 値が最大となり，閾値を 3.6 から 4.6 まで変化させた場合の再現率，適合率の平均値を採用した．

手法 B は，構文情報のみを考慮した動詞連体の省略可能性判定手法を仮定し，その実行結果を得る．本手法で動的に省略可能である動詞連体を認定可能であるが，それが認定性能に貢献しているかを調べるために，手法 A と比較する対象として行なう．

手法 C は，文脈補正項を導入しない手法である．本手法では，対象の動詞連体の内容および前後の文脈を考慮した文脈補正項を導入している．この文脈補正項が認定性能に貢献しているのかを調べるため，手法 A と手法 C を比較し，文脈補正項の影響を調べる．ただし，手法 C の場合は，文脈補正項による重みの上昇がないので，手法 A の再現率と等しい再現率になるように閾値を調整して適合率を算出した．実験結果を表 2 に示す．

5 考察

評価の結果，提案手法は再現率 72.7%，適合率 79.3% と比較的良好な結果を得ることができたと考える．構文情報のみを考慮した手法 B と比較すると，再現率が 17.7% 低下したのに対して適合率は 30.0% 上昇している．構文情報のみの手法 B よりも，名詞の重要度や修飾されやすさ，文脈を考慮に入れた本手法が，高い精度で省略可能な動詞連体を認定することができた．よって本手法は有効であると考える．また，文脈補正項を導入しなかった手法 C と比較してみても，手法 A は同じ再現率で適合率が 13.9% 上昇しており，本手法で示した文脈補正項が妥当であると考える．

固有名詞や複合名詞などの，その名詞だけで意味が分かる名詞を修飾する動詞連体が，一般知識等から補完がしやすいため省略可能と認定された．また，文脈から補完可能な情報であるため，省略可能と認定された動詞連体も存在した．正解例をいくつか示す．なお，

下線で示された部分が省略可能と認定された動詞連体である.

- 例 1: 簡単に録画予約ができる Gコード内蔵型のVTRを投入しているほか、ビデオカセットを近付けるとカセットの挿入口が自動的に開くなど消費者の目を引く機能も盛り込んでいる。
- 例 2: 一方アジア・太平洋地域は海面水位上昇の影響を受けやすい低海岸国が多いにもかかわらず、これまで十分な観測ができず基礎的な情報の蓄積も遅れていた。

例1は「簡単に録画予約ができる」が「Gコード内蔵」を修飾している. これは一般知識から補完できる情報であるので省略可能であると考える. 修飾先の名詞が「Gコード内蔵」という複合名詞であり, 修飾頻度が低く, 式 (1) におけるエントロピー $E(l(n))$ が小さくなった. (「内蔵」のエントロピーは 0.693 である. それに対し, 例えば「案」のエントロピーは 4.13 である.) そして,「Gコード内蔵」という複合名詞が全コーパス中に出現する頻度は 1 であり, 式 (1) における $idf(n)$ の値が大きくなった. よって, 動詞連体の重みが小さくなり省略可能と認定された. 例3は省略箇所の情報がないと意味が分からない文になるが, 文脈から補完できる情報なので省略可能と認定された. 例3の場合は, 文書の第一文に「海面の上昇など生活環境に悪影響を及ぼす地球温暖化を防止するため...」という表現があり,「低海岸国」に対する修飾表現「海面水位上昇の影響を受けやすい」は, 説明されなくても補完できる内容である.

次に不正解例をいくつか示す.

- 例 3: 松下電器産業と松下電子工業は共同で、世界最高速で動作する １６メガビットＤＲＡＭを開発.
- 例 4: ドイツではこれまで何度か、高速道路有料化案が出たが、国内、国外の反対でつぶれてきた いきさつがある。

例3の場合は「１６メガビットＤＲＡＭ」を修飾している「世界最高速で動作する」が省略可能と認定されてしまった. この文において「世界最高速で動作する」が最も重要な部分である. そのため, 情報欠落が大きいので不正解と認定された.「ＤＲＡＭ」のエントロピーは 2.16 と比較的低く, 特に, 修飾先の複合名詞「１６メガビットＤＲＡＭ」に含まれる名詞の数 $J(n)$ が 3 であるため, 動詞連体の重みが小さくなり, 省略可能と認定されてしまった. 例4の場合は,「いきさつ」を修飾している動詞連体が省略可能と認定されてしまった.「いきさつ」は連体修飾表現がないと意味が分からない名詞であるが, 全コーパスにおける頻度が少ないため, 重要な名詞であると認識される. よって, 重みの値が小さくなり「いきさつ」を修飾している動詞連体が省略可能であると認定された. もし,「いきさつ」と同じ意味である「経緯」であったなら, コーパスにおける頻度も高く, エントロピーも高いため省略可能と認定されなかった. この問題に対しては, エントロピーが高く, コーパスにおける頻度が高いような, 類似した意味をもつ名詞が存在したならば, 省略可能と認定しないといった制限を設けることで対処できると考える.

6 結び

本論文では, 一般のコーパスから, 省略可能な動詞連体修飾表現を自動獲得し, 獲得した言語知識を用いて要約を行なう手法を提案した. 具体的には, 省略できる可能性のある動詞連体が修飾している名詞に対して, "修飾されやすさ", "修飾多様性" をコーパスから調べ, 修飾される頻度が低い, もしくは, 修飾する動詞の種類が限定されている名詞に係る動詞連体を省略可能と認定する. 同時に, その名詞の重要度を算出し, 重要な名詞に係る動詞連体を省略可能と認定する. また, 動詞連体の内容および前後の文脈を考慮して, その動詞連体に含まれている名詞が以前の文にも含まれている名詞である場合には, 省略可能と認定されやすくなる. 逆に, 重要な情報が含まれている場合には省略可能と認定されにくくなるような工夫を行なっている. 評価実験によって, 本手法による省略可能な動詞連体は再現率 72.7%, 適合率 79.3%を示し, 比較的, 良好な結果であった. よって, 本手法は有効であると考える. 今後の課題として, 修飾先の名詞がなんらかの連体修飾表現をともなわないと意味をなさない場合, それを判別することで精度を高めるような工夫を行なうことが挙げられる.

謝辞

言語データとして, 日本経済新聞 CD-ROM 版の使用を許可して頂いた日本経済新聞社に深謝する.

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. 5, pp. 1–25 (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] 大竹清敬, 岡本大吾, 児玉充, 増山繁: 重要文抽出, 自由作成要約に対応した新聞記事要約システム YELLOW, 情報処理学会論文誌データベース, Vol. 43, No. SIG2(TOD13) 掲載予定 (2002).

[4] 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, Vol. 97-NL-122, No. 13, pp. 83–89 (1997).

[5] 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).

[6] 加藤直人, 浦谷則好: 局所的要約知識の自動獲得手法, 自然言語処理, Vol. 6, No. 7, pp. 73–92 (1999).

[7] Salton, G.: *Automatic Text Processing*, Addison Wesley (1988).