自然言語処理　150－26
(2002. 7. 16)

省略可能な2重修飾表現の教師なし学習による自動獲得

酒井　浩之†　　増山　繁†

† 豊橋技術科学大学 知識情報工学系　〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘1-1
E-mail: †sakai@smlab.tutkie.tut.ac.jp, ††masuyama@tutkie.tut.ac.jp

あらまし　文内要約の一要素技術として，省略可能な多重(連用・連体)修飾表現を，教師なしで自動的に獲得する手法を提案する．具体的には，(連用・連体)修飾表現と修飾先の名詞や動詞との結合度を相互情報量を元にした計算式で重みとして計算する．そして，複数の修飾表現の重みの相対的な差で省略可能な修飾表現を認定する．すなわち，重みが最も大きい修飾表現を省略不可とし，それ以外の修飾表現は最も重みの大きい修飾表現との重みの差が大きいとき省略可能と認定する．評価実験によって，本手法による省略可能な多重連体修飾表現は，精度73.4%，再現率45.7%，多重連用修飾表現は，精度74.7%，再現率42.9%であり，比較的，良好な精度を示した．
キーワード　多重修飾表現の省略，コーパス，テキスト自動要約．

Unsupervised Acquisition about the abbreviation possibility of double(multiplex) verb/noun phrases

Hiroyuki SAKAI† and Shigeru MASUYAMA†

† Department of Knowledge-based Information Engineering, Toyohashi University of Technology　1-1 Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi 441-8580, Japan
E-mail: †sakai@smlab.tutkie.tut.ac.jp, ††masuyama@tutkie.tut.ac.jp

Abstract　This paper proposes an unsupervised method of acquiring knowledge about the abbreviation possibility of multiplex verb/noun phrases. Our method calculates weight values of verb/noun phrases and the weight values are calculated by mutual information based on degree of combination of verb/noun phrases and modified verbs/nouns. And, our method recognizes verb/noun phrases possible to be abbreviated by comparing weight values of verb/noun phrases. The evaluation of our method by experiments shows that the precision attains 73.4% and the recall attains 45.9% in case of noun phrases and the precision attains 74.7% and the recall attains 42.9% in case of verb phrases.
Key words　abbreviation of multiplex verb/noun phrases, corpus, summarization.

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[3]．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[8]．本稿では，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，文中の省略可能な多重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．ここで，多重修飾表現とは，ある一つの動詞に，複数の連用修飾表現が係っている場合，もしくは，ある一つの名詞に，複数の連体修飾表現が係っている場合を指す．不要箇所省略による要約手法として，大竹ら[4]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省略を提案している．若尾ら[7]や山崎ら[9]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[2]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，文書群の統計的情報を用いて自動的に獲得し，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．

文書群の統計的情報を利用した省略可能な修飾表現の研究として，我々は動詞連体修飾表現に着目した研究を行なった[5]．文献[5]では，動詞連体修飾表現において，修飾される名詞の“修飾されやすさ”，“修飾多様性”を，名詞とそれを修飾する動詞の種類や修飾頻度から調べ，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．ただし，文献[5]では，動詞連体修飾表現のみに対応しており，その他の連体修飾表現には対応していない．しかし，一般の連体修飾表現の省略可能性を認定するのは，連体修飾表現の多様性から困難であり，例えば動詞連体のような何らかの文構造に限定する必要があると考える．そこで，本稿では不要箇所の単位として多重修飾表現に着目する．

多重修飾表現の省略は大竹ら[4]が行なったが，大竹らの着目した多重修飾表現は，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，人手で作成した規則があてはまる用例にのみ適用でき，網羅性が保証されていないという問題がある．また，一般には，多重修飾表現は連体修飾表現よりも連用修飾表現に多く見られる現象である．そのため，大竹らのアプローチで動詞を修飾している省略可能な連用修飾表現を認定するには，膨大な要約規則を人手で新たに作成する必要があり，人的コストが膨大となる．それに対して，本論文で提案する手法ではコーパスの統計的情報を使用して，修飾表現と修飾先の名詞や動詞との結合度を相互情報量を元にした計算式で計算し，その値で省略可能な修飾表現を認定する．そのため，文書群の統計的情報を考慮にいれた，動的な省略可能な多重修飾表現の認定を行なうことができる．

我々は，多重連用修飾表現において，文書群の統計的情報を利用した省略可能な修飾表現の認定を行なう手法を既に提案した[6]．文献[6]では，省略できる可能性のある連用修飾表現を含む節に対して，同一の動詞をもち，かつ，格助詞出現の差異が認められる節をコーパスから検索し，検索された節対から省略可能な連用修飾表現を認定する．しかし，文献[6]の手法は，多重連用修飾表現のみに対応しており，多重連体修飾表現には対応していない．それに対して，本稿で提案する手法は，連体修飾表現，連用修飾表現を含む全ての多重修飾表現に同一の手法で対応できる．また，文献[6]は，省略認定のために同一の動詞をもち，かつ，格助詞出現の差異が認められる節をコーパスから検索する必要があり，前処理を含めた全体の処理時間が高いが，本手法では，単語の頻度情報のみで認定ができるので，文献[6]の手法と比べて前処理を含めた全体の処理時間が低いという利点もある．

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多重修飾表現を示す．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

大竹ら[4]らは要約規則を人手で作成し，2重修飾表現を省略したが，連体修飾表現における2重修飾表現に限定していた．しかし，実際は，連用修飾表現における多重修飾表現の方が出現頻度が大きい．また，3重修飾表現のような，2重修飾表現よりも多重度の高い多重修飾表現も存在する．大竹らのアプローチでは，出現頻度の大きく多様性の大きい連用修飾表現に対しては対応できない．その理由は，大竹らのアプローチでは，人手によって要約規則を作成する必要があるため，出現頻度の大きく多様性の大きい多重連用修飾表現に対応するためには，多くの要約規則を作成する必要があるためである．しかし，人手による作業量には限界があり，多くの事例を網羅するような要約規則をそろえるのは困難であると考える．それに対して，本稿で提案する手法は，対象としたコーパスの統計的情報に基づいて，省略可能な多重修飾表現を認定する．そのため，人手による要約規則作成は不要であり，多重連体修飾表現のみならず，多重連用修飾表現にも対応可能である．そして，修飾表現の多重度に関しては制限がない．具体的には，複数の修飾表現と，修飾先の名詞や動詞との結合度を相互情報量を元にした計算式で重みを計算する．そして，複数の修飾表現の重みの相対的な差で，省略可能な修飾表現を認定する．すなわち，重みが最も大きい修飾表現を省略不可とし，それ以外の修飾表現は最も重みの大きい修飾表現との重みの差が大きいとき，省略可能と認定する．

2.1 多重連用修飾表現の省略

多重連用修飾表現における省略可能な連用修飾表現を認定する手法について述べる．ここで，多重連用修飾表現とは，ある動詞に対して，複数の連用修飾表現が係っている事例のことである．例えば，「ＥＣは市場統合に続いて通貨・政治統合を目指している.」は動詞「目指す」に対して，「ＥＣは」，「市場統合に続いて」，「通貨・政治統合を」の3つの連用修飾表現が係っている．このような事例のことを指す．

手法の説明の前に，以下のように連用修飾表現を記号で定義する．

$E(M, c_j, V)$: ある動詞Vに対して，品詞(もしくは意味素性)Mが，係り関係子c_jで係っている連用修飾表現，

Mとは，連用修飾表現において動詞に係っている語を表す．

表 1　その他の主な係り関係子 (連用修飾表現)

M の品詞	係り関係子
名詞	ガ格，ヲ格，デ格，ニ格
名詞	無格，隣接
名詞	～で判，～を～に
動詞	複合辞連用，～と
形容詞	弱連用，～く
指示詞	マデ格，デ格

ただし,「ガ」「ヲ」などの助詞は除く．また，M が名詞であった場合は，M は，その名詞の意味素性となる．ここで，意味素性として日本語語彙大系 [1] の単語体系を利用した．日本語語彙大系 [1] の単語体系の知識を用いることにより，辞書に記載されている名詞に対して意味素性を割り当てることができる．ただし，意味素性が存在しない名詞は，その名詞自身とする．例えば,「ＥＣは市場統合に続いて通貨・政治統合を目指している.」で，動詞「目指す」に対する連用修飾表現「ＥＣは」では，名詞「ＥＣ」が動詞に係っているので，M は名詞「ＥＣ」となる．なぜなら，名詞「ＥＣ」の意味素性は存在しないからである.「通貨・政治統合を」では，名詞「統合」が動詞に係っているので M は「統合」の意味素性である「2229 結合」となる.「市場統合に続いて」では,「続く」が動詞に係っているので M は「続く」となる．

次に係り関係子 c_j について説明する．これは，M が動詞 V に係る際の係り関係を表す．なお，係り関係の表記に関しては，構文解析器として使用した KNP[(注1)] version2.0b6 に準拠する．具体的には，KNP 解析結果の「<係:　>」タグと,「<ID:（　）>」タグを利用する．例えば上記の例で「通貨・政治統合を」はヲ格で動詞に係ってるので，c_j = ヲ格 とする．格助詞の伴わない「市場統合に続いて」の場合は，c_j = 複合辞連用 とする．表 1 に，連用修飾表現における，その他の主な係り関係子 c_j を示す．以下に手法を説明する．

省略可能な多重連用修飾表現を認定するアルゴリズム

Step 1　ある動詞 V に連用修飾表現 $E(M,c_j,V)$ が係っているとする．ここで，各連用修飾表現 $E(M,c_j,V)$ に対して，重み $W(E(M,c_j,V))$ を計算する．

$$W(E(M,c_j,V)) = \frac{A(M)\times P(E(M,c_j,V))}{f(M,c_j,V)} \times \frac{1}{\exp(I(M,c_j,V))} \quad (1)$$

$$I(M,c_j,V) = \log(S\times\frac{f(M,c_j,V)}{f(M,c_j)\times f(c_j,V)}) \quad (2)$$

$f(M,c_j,V)$:　語 (意味素性)M が係り関係子 c_j によって動詞 V を修飾する頻度,

$f(M,c_j)$:　語 (意味素性)M が係り関係子 c_j によって動詞を修飾する頻度

$f(V,c_j)$:　係り関係子 c_j によって動詞 V を修飾する頻度,

S:　全コーパスにおける多重連用修飾表現の頻度,

(注1)：http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html

表 2　多重修飾表現における品詞の頻度

M の品詞	頻度 (連用修飾表現)	頻度 (連体修飾表現)
名詞	1277730	77754
動詞	178203	22217
形容詞	53918	12709
副詞	70479	2243
指示詞	15763	4086
接続詞	19088	108
S	1615181	119117

$A(M)$:　M の品詞の総数，具体的には，名詞，動詞，形容詞，接続詞，副詞，指示詞の 6 つである．

$P(E(M,c_j,V))$:　連用修飾表現 $P(E(M,c_j,V))$ の文節の数,

Step 2　$W(E(c_j,V))$ を，同じ文の動詞 V に係っているいくつかの連用修飾表現の重みの最大値で正規化する．ここで，文の動詞 V には k 個の連用修飾表現が係っているものとする．

$$Ws(E(M,c_j,V)) = \frac{W(E(M,c_j,V))}{\max_{i=1,2,\ldots,k} W(E(M,c_i,V))} \quad (3)$$

Step 3　$Ws(E(M,c_j,V))$ が，ある閾値以下の連用修飾表現を省略可能と認定する．ただし，c_j がガ格，ヲ格，ニ格，ト格，～と，であった場合は無条件に省略不可能と認定する．

□

Step 2 の処理によって，多重連用修飾表現の重みを最大値で正規化しているので，複数ある $Ws(E(M,c_i,V))$ のどれか 1 つは値が 1 になっており，必ず省略不可能と認定される．

式 (2) の $I(M,c,V)$ は，語 (意味素性)M と係り関係子 c_j によって構成されている連用修飾表現と，係り関係子 c_j によって修飾されている動詞 V との相互情報量である．連用修飾表現の重み $W(E(M,c_j,V))$ は，この相互情報量が大きいほど小さい値をとる．そして，連用修飾表現の重みが小さいほど省略可能と認定されやすくなる．これは，相互情報量が大きい連用修飾表現と動詞の組合せは一般にコーパス中に多数存在し，係り先の動詞から連想できるので，省略可能であるという仮定に基づく．

$A(M)$ は，動詞を修飾している語 (意味素性)M の品詞の総頻度である．基本的に名詞が多いが，次に動詞が多い．接続詞や指示詞は少ない．これは，語 (意味素性)M が少ない方が，省略可能と認定できる場合が多いという経験に基づいて，計算式に導入した．よって，連用修飾表現の中でも $A(M)$ の小さい接続詞，形容詞，指示詞，副詞は省略可能となりやすい．また，$A(M)$ は修飾先が動詞か名詞かで大きく変化し，それにともない，修飾表現の重みが変化するため，省略可能な修飾表現も変化する．表 2 に日経新聞 93 年 1 月 1 日～6 月 30 日までを対象とした場合の各品詞の頻度 $A(M)$ を示す．また，多重修飾表現の総頻度 S を示す．表 2 から，多重連用修飾表現のほうが，多重連体修飾表現に比べてはるかに出現頻度が大きいことが分かる．また，各品詞の出現頻度から連用修飾表現では名詞以外が省略可能となりやすく，連体修飾表現では名詞と動詞以外が省略可能となりやすくなる．なお，連用修飾表現，連体修飾表現，

表 3　その他の主な係り関係子 (連体修飾表現)

M の品詞	係り関係子
名詞	ノ格，ト格，隣接
動詞	動詞連体，複合辞連体，
形容詞	形判連体，弱連体

どちらの場合でも，M の品詞が名詞である場合が他の場合に比べて極端に多いため，必然的に，名詞以外の修飾表現が省略可能となりやすくなる．しかし，M の品詞が名詞である場合にも省略可能な修飾表現は存在するため，M の品詞が名詞である場合は $A(M)$ の値を $A(M)/2$ とした．

$P(E(M,c_j,V))$ は，連用修飾表現 $P(E(M,c_j,V))$ の文節の数である．例えば，「ＥＣは市場統合に続いて通貨・政治統合を目指している.」の連用修飾表現「市場統合に続いて」では，$P(E(M,c_j,V))=2$ となる．これは，$P(E(M,c_j,V))$ の値が大きくなるような長い連用修飾表現を省略しにくくするため，計算式に導入した．なお，$P(E(M,c_j,V))$ が 4 以上の連用修飾表現は，その重みの大きさにかかわらず無条件に省略不可とした．

その他の制約として，係り先の動詞が「する」「なる」「ある」の場合は，それを修飾している連用修飾表現を無条件に省略不可とした．

2.2　多重連体修飾表現の省略

多重連用修飾表現における省略可能な連体修飾表現を認定する手法について述べる．ここで，多重連体修飾表現とは，ある名詞に対して，複数の連体修飾表現が係っている事例のことである．例えば,「政権を脅かす有力なライバルが存在しない」という文では，名詞「ライバル」に,「政権を脅かす」「有力な」という 2 つの連体修飾表現が係っている．このような事例のことである．

手法の説明の前に，以下のように連体修飾表現を記号で定義する．

$E(M,c_j,N)$　語 (意味素性)M が，係り関係子 c_j によって名詞 N を修飾している連体修飾表現,

表 3 に，連体修飾表現における，その他の主な係り関係子 c_j を示す．以下に手法を説明する．

省略可能な多重連体修飾表現を認定するアルゴリズム

Step 1　ある名詞 N に連体修飾表現 $E(M,c_j,N)$ が係っているとする．ここで，各連体修飾表現 $E(M,c_j,N)$ に対して，重み $W(E(M,c_j,N))$ を計算する．

$$W(E(M,c_j,N)) = \frac{A(M)\times P(E(M,c_j,N))}{f(M,c_j,N)} \times \frac{1}{\exp(I(M,c_j,N))} \quad (4)$$

$$I(M,c_j,N) = \log(S\times\frac{f(M,c_j,N)}{f(M,c_j)\times f(c_j,N)}) \quad (5)$$

$f(M,c_j,N)$:　語 (意味素性)M が係り関係子 c_j によって名詞 N を修飾する頻度,

$f(M,c_j)$:　語 (意味素性)M が係り関係子 c_j によって名詞を修飾する頻度

$f(N,c_j)$:　係り関係子 c_j によって名詞 N を修飾する頻度,

S:　全コーパスにおける多重連体修飾表現の頻度,

$A(M)$:　M の品詞の総数，具体的には，名詞，動詞，形容詞，接続詞，副詞，指示詞の 6 つである．

$P(E(M,c_j,V))$:　連体修飾表現 $P(E(M,c_j,N))$ の文節の数，4 以上で無条件に省略不可

Step 2　$W(E(c_j,N))$ を，同じ文の名詞 N に係っているいくつかの連体修飾表現の重みの最大値で正規化する．ここで，文の名詞 N には k 個の連体修飾表現が係っているものとする．

$$Ws(E(M,c_j,N)) = \frac{W(E(M,c_j,N))}{\max_{i=1,2,\cdots,k} W(E(M,c_i,N))} \quad (6)$$

Step 3　$Ws(E(M,c_j,N))$ が，ある閾値以下の連体修飾表現を省略可能と認定する．ただし，格がガ格，ヲ格，ノ格，ト格，隣接，～と，未格であった場合は無条件に省略不可能とする．

□

その他の制約として，係り先の名詞が「こと」「もの」の場合は，それを修飾している連体修飾表現を無条件に省略不可とした．

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，66,686 記事を採用した．形態素解析器として JUMAN (注2) version 3.5 を，構文解析器として KNP version2.0b6 を採用した．本手法によって，仮に閾値 0.24 とした場合の省略可能となった多重連用修飾表現，多重連体修飾表現をいくつか示す．例において，下線の修飾表現が省略可能と認定された多重修飾表現である．

- 市場統合はＥＣが八五年から 八年越しで 進めてきた計画
- 総額五兆円に達する不良債権を 融資金融機関の支援で 分離、今月に発足する共同債権買取機構などで 段階的に 処理を進める。
- 政権を脅かす 有力なライバルが存在しないことに加え、政権党のメリットを生かし、行政を味方に引き入れ、利益還元政治を実現できたことが大きい。

4.　評価実験

実装した提案手法の評価を行なった．評価方法は，対象記事群から省略可能な多重修飾表現の正解データを作成し，精度，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，省略しても妥当な多重修飾表現を人手で抽出し，作成した．再現率，精度の定義を示す．

$$\text{再現率} = \frac{\text{本手法による結果と正解データとが一致する数}}{\text{正解データの省略可能な修飾表現の数}}$$

$$\text{精度} = \frac{\text{本手法による結果と正解データとが一致する数}}{\text{本手法によって省略可能と判定された修飾表現の数}}$$

4.1　多重連体修飾表現における省略認定の評価

多重連体修飾表現による省略認定には先行研究がある．それは，大竹ら [4] らの行なった，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の省略手法である．大竹らは連体

(注2)：http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

表 4　実験結果 (Yellow との比較)

手法	精度 (%)	再現率 (%)	省略個数
提案手法	71.2	49.1	73
Yellow	65.8	49.1	79

表 5　実験結果 (連体修飾表現)

閾値	精度 (%)	再現率 (%)	省略個数
0.12	75.4	40.6	57
0.14	76.7	43.4	60
0.16	76.7	43.4	60
0.18	74.6	44.3	63
0.2	75.0	45.3	64
0.22	73.1	46.2	67
0.24	72.5	47.2	69
0.26	70.4	47.2	71
0.28	70.8	48.1	72
0.3	70.8	48.1	72
0.32	71.2	49.1	73
平均	73.4	45.7	66.2

の二重修飾表現において，36 個の規則を人手により作成し，省略可能な連体修飾表現の認定を行なった．(以降，大竹らの手法を，それに基づいて開発されたテキスト自動要約システムの名称から Yellow と呼ぶ．) そこで，我々は，Yellow を実装し，本手法の比較対象として評価実験を行なった．対象記事は，日経新聞記事 1 月 1 日から 6 月 30 日まで，66,686 記事とし，この記事群を用いて，本手法に必要な頻度情報などの統計情報を入手した．また，対象記事から 157 記事を無作為に選び，この中に含まれる多重連体修飾表現を全て抽出し，省略しても妥当な連体修飾表現を人手で抽出して正解データを作成した．なお，157 記事の中には 145 の多重連体修飾表現の事例が存在した．本手法と Yellow で，正解データにおける精度，再現率を算出した．その実験結果を表 4 に示す．なお，本手法は閾値によって精度，再現率が変化する．それに対して，Yellow では固定である．そこで，再現率が Yellow と同じ値になるように閾値を設定し，精度を算出する．次に，本手法において，閾値を 0.12〜0.32 まで 0.02 ずつ変化した場合の精度，再現率を表 5 に示す．

4.2　多重連用修飾表現における省略認定の評価

連体修飾表現の評価実験と同じく，対象記事は，日経新聞記事 1 月 1 日から 6 月 30 日まで，66,686 記事とし，この記事群を用いて，本手法に必要な頻度情報などの統計情報を入手した．ただし，多重連用修飾表現は多重連体修飾表現と比べて出現頻度が大きいので，評価実験における記事数は 12 記事とした．対象記事から 12 記事を無作為に選び，この中に含まれる多重連用修飾表現を全て抽出し，省略しても妥当な連用修飾表現を人手で抽出して正解データを作成した．なお，12 記事の中には 199 の多重連用修飾表現の事例が存在した．本手法において，閾値を 0.12〜0.32 まで 0.02 ずつ変化した場合の精度，再現率を表 6 に示す．

表 6　実験結果 (連用修飾表現)

閾値	精度 (%)	再現率 (%)	省略個数
0.12	75.0	39.0	72
0.14	76.0	41.2	75
0.16	75.0	41.2	76
0.18	75.3	41.9	77
0.2	75.6	42.6	78
0.22	74.7	42.6	79
0.24	75.3	44.0	81
0.26	73.5	44.0	83
0.28	73.5	44.0	83
0.3	74.1	45.5	85
0.32	74.1	45.5	85
平均	74.7	42.9	79.5

5.　考　　察

Yellow との比較であるが，再現率が同一になるように閾値を調節した場合の精度は，Yellow が精度 65.8%であるのに対して提案手法は 71.2%であり，約 6%の向上があった．Yellow は，二重連体修飾表現において，36 個の要約規則を人手により作成し，省略可能な連体修飾表現の認定を行なっている．しかし，36 個の要約規則を作成するのは膨大な人的コストが必要と考える．本手法は，このような人的コストを必要とせず，Yellow を超える精度を達成することができた．

Yellow の省略認定が失敗した事例は，連体修飾表現に含まれている文節の数が多い，長い連体修飾表現である場合があった．これは，Yellow の手法が文構造のみを考慮した要約規則に基づき，連体修飾表現と修飾先の名詞との結合度や，連体修飾表現の長さといった点を考慮していないためと考える．本手法は，連体修飾表現と名詞との相互情報量の他にも，文節の数が多い，長い連体修飾表現は省略不可とするため，そのような連体修飾表現を省略可能と認定しなかった．よって，Yellow より優れた精度を得ることができたと考える．

本手法は閾値によって精度と再現率を変化させることができる．表 5 や表 6 のように，基本的に閾値を下げるほど，精度が高くなるが再現率も低くなる．表 5 や表 6 に示すように，閾値を 0.12〜0.32 まで 0.02 ずつ変化した場合の平均精度は，連体修飾表現の場合は 73.4%，連用修飾表現の場合は 74.7%であった．よって，人手による要約規則による手法の精度が 65.8%であることを考慮すると，精度よく省略可能な修飾表現を認定できたと考える．

しかし，連体修飾表現，連用修飾表現，ともに全体的に再現率が低い．これは閾値を上げることで，ある程度，解決できるが，精度が下がってしまう．そのため，新たに再現率を上げる方法を工夫する必要がある．例えば，「便利な」，「理想的な」といった連体修飾表現は省略可能である場合が多い．しかし，本手法では，複数ある修飾表現の重みの相対的な差で省略可能かどうかを認定するため，このような修飾表現も省略不可とする場合がある．そこで，大規模コーパスに対して本手法を適用して省略可能な修飾表現を抽出し，省略可能と認定される頻度の

大きい，定型的な修飾表現を抽出する．そして，そのような定型的な修飾表現を無条件に省略可能と認定すれば，それほど精度を落さずに再現率が向上すると考える．この方法によって抽出される知識は，要約規則と呼べるものであるが，それを人手ではなく，精度よく自動で抽出できるところが従来手法と異なると考える．

6. 結　び

本論文では，一般のコーパスから，省略可能な多重修飾表現を認定し，要約を行なう手法を提案した．要約を行なう文書群の統計情報を用いて省略可能な多重修飾表現の認定を行なうので，従来の要約規則を人手で作成する手法よりも，人的コスト，網羅性，精度の点で有利である．具体的には，複数の修飾表現と，修飾先の名詞や動詞との結合度を相互情報量を元にした計算式で重みを計算する．そして，複数の修飾表現の重みの相対的な差で，省略可能な修飾表現を認定する．すなわち，重みが最も大きい修飾表現を省略不可とし，それ以外の修飾表現は最も重みの大きい修飾表現と重みの差が大きいとき，省略可能と認定する．

評価実験によって，本手法による省略可能な多重連体修飾表現は，精度 73.4%，再現率 45.7%，多重連用修飾表現は，精度 74.7%，再現率 42.9%であり，比較的，良好な精度を示した．

また，従来の人手による要約規則作成による手法と比較した結果，同一の再現率において，約 6%の精度向上を示した．これは，要約規則を作成する膨大な人的コストを必要とせず，文内要約に必要な省略可能な修飾表現を認定できたと考える．

今後の課題として，精度を落さずに再現率の向上を達成することと，文献 [6] の手法との性能比較を挙げる．

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．また，日本語語彙大系から意味分類を取得するために用いた形態素解析システム ALT-JAWS ver.2.0. の使用を許可して頂いた日本電信電話 (株) に深謝する

文　献

[1] 池原悟, 宮崎正弘, 白井論, 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, 岩波書店 (1997).
[2] 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).
[3] 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).
[4] 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), pp. 37–47 (2002).
[5] 酒井浩之, 増山繁: 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, 言語処理学会第 8 回年次大会発表論文集, pp. 627–630 (2002).
[6] 酒井浩之, 篠原直嗣, 増山繁, 山本和英: 連用修飾表現の省略可能性に関する知識の獲得, 自然言語処理, Vol. 9, No. 3 (2002).
[7] 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).
[8] 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).
[9] 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).