略語とその原型語との対応関係のコーパスからの自動獲得手法の改良

酒井　浩之[†]　増山　繁[†,††]

略語とその略語に対する元の語 (原型語と定義) との対応関係を，コーパスから自動的に獲得する手法を提案する．本手法は，同一の再現率においてより高い精度を達成できるように，我々の既提案手法 (酒井浩之 増山繁 2002) を改良したものである．このような知識は，情報検索や文書要約などにおいて有用である．本手法は，まず，略語候補とそれに対応した原型語の候補を，それらを構成している文字情報から獲得する．そして，略語候補と原型語の候補の名詞間類似度を計算することで，略語とその原型語との対応関係を取得する．例えば，略語「原発」に対して，原型語「原子力発電所」のような対応関係を取得できる．なお、本手法はコーパスに出現する各名詞が略語か原型語であるかどうかの情報が与えられていることを前提としていない。評価の結果，名詞間類似度の閾値を 0.4 に設定した場合，精度 73.4% の結果を得た．本手法と既提案手法とを比較した結果，同一の再現率においてより高い精度を達成し，既提案手法よりも有効な手法であることを確認した．

キーワード:　略語原型語対応獲得，コーパスからの知識獲得，換言

Improvement of the Method for Acquiring Knowledge from a Single Corpus on Correspondences between Abbreviations and Their Original words

Hiroyuki Sakai[†] and Shigeru Masuyama[†,††]

We propose a method for acquiring knowledge from a single corpus on correspondences between abbreviations and their original words. This is an improvement of our previous method so that higher precision is attained for the same recall. This knowledge is useful for such tasks as information retrieval, word sense disambiguation and summarization. Our method searches “abbreviation candidates” and “original word candidates” corresponding to the abbreviation candidates by using information of characters composing them. Then, in order to decide a correspondence between an abbreviation and its original word, the similarity between the abbreviation candidate and the original word candidate is calculated by using statistical information in the single corpus. For example, a correspondence between abbreviation “原発 *gempatsu* (a nuclear power station)” and original word “原子力発電所 *genshiryoku hatsudensho* (a nuclear power station)” is extracted by our method. Here, our method does not presume that information whether each noun in the corpus is an abbreviation or an original word is given. Experimental results show that our method is promising, as the precision attains 73.4%. We compare our method with our previous method and experimental results suggest that our method is able to extract correspondences between abbreviations and original words more appropriately than our previous method.

KeyWords: *Extraction of coprrespondences between abbreviaions and their original words, Knowledge acquisition from a corpus, Paraphrasing*

1　はじめに

日本語には，ある名詞から一部の文字を省略することで意味内容を保ったまま別の表現に変換する事例が多く存在し，一般に略語と定義される．例えば，「原発」は，「原子力発電所」の略語であり，「原子力発電所」からいくつかの文字が省略されているにもかかわらず，同一の意味を持つ．また，複合名詞から一部の名詞が省略されても意味が変わらない，あるいは，読み手が省略された名詞を推定できる場合が存在する．例えば「バブル経済崩壊」から「経済」が省略されて「バブル崩壊」と表記されても，一般知識から「経済」を推定することができる．このような名詞と略語との対応も要約や検索において有用な知識であるため，抽出対象とする．そのため，本稿では，複合名詞から名詞が省略された場合も同じく略語と定義し，略語とそれに対応した元の語との対応関係をコーパスから自動的に獲得する手法を提案する．(以降，略語に対応した元の語を原型語と定義する．)

略語とその原型語との対応関係を取得できれば，情報検索や文書要約において有用である．

- 検索：検索キーワードとして略語を入力した場合でも，その略語に対応する原型語を含む文書も検索できるようになる．新しい略語は常に発生し続けるため，人手による登録では多大な労力が必要となる．本手法はコーパスに含まれている略語とその原型語との対応を自動的に取得できるので，インターネットのニュース記事等をコーパスに使用すれば，最新の略語と原型語との対応も獲得可能である．
- 要約：名詞の略語は元の名詞から一部の文字が削除された表現であるので，文字数が少なくなるにもかかわらず同一の意味をもつ名詞である．そのため，文中の名詞を略語に置き換えることによる文内要約に利用できる．また，複数文書要約においては異った文書中で略語とその原型語が混在している場合があり，冗長性の削除のために略語とその原型語を認識できる必要がある．

略語とその原型語との対応関係の獲得は言い換え知識獲得の一種である．なぜなら，略語と原型語は表記が異なるにもかかわらず同じ意味を持つからである．

言い換え知識獲得は，例えばシソーラスの構築や語のクラスタリングといった研究(例えば，(Hindle 1990) (Pereira, Tishby, and Lee 1993) (Lin 1998) (真田亜希子, 舟宝貴志, 梅村恭司, 山本英子 2003))とは同じ意味を持つ表現を獲得しなければならないという点で異なる．言い換え知識獲得の研究としてさまざまな手法が提案されている(乾健太郎 藤田篤 2004)．例えば，原著に対する複数の翻訳本から言い換え知識を獲得する手法(Barzilay and Mckeown 2001)，「サ変名詞+する」から動詞相当句へ言い換える手法(近藤恵子, 佐藤理史, 奥村学 1999)，「AがVするB」のような動詞連体修飾表現を「AのB」へ言い換える手法(片岡明, 増山繁, 山本和英 2000)，

† 豊橋技術科学大学知識情報工学系, Department of Knowledge-based Information Engineering, Toyohashi University of Technology
†† 豊橋技術科学大学インテリジェントセンシングシステムリサーチセンター, Intelligent Sensing System Research Center, Toyohashi University of Technology

などがある．本手法は略語と原型語との対応関係をコーパスから獲得する手法であり，それに対応した手法を開発する必要がある．

関連研究として，括弧表現から語の対応を自動抽出する手法が提案されている (Hisamitsu and Niwa 2001)．例えば「朝鮮民主主義人民共和国（北朝鮮）」から「朝鮮民主主義人民共和国」と「北朝鮮」との語の対応を獲得している．この手法により本稿で述べる手法では獲得できない対応を獲得することができるが，コーパスに括弧表現として記載されている対応しか獲得できない．例えば，コーパスに「原子力発電所 (原発)」という文字列が出現しないと，「原子力発電所」とその略語「原発」との対応は獲得できない．しかし，その略語が一般的に知られている表現であれば，名詞 (略語) の表現はコーパス中に出現しないと考える．それに対して，本手法では括弧表現のような特別な文構造を必要としないため，「原子力発電所」とその略語「原発」との対応を獲得できる．

英語の略語復元に関する関連研究として，プログラムソースに出現する略語を規則に基づいて復元する手法が提案されている (Rowe and Laitinen 1995)．日本語の略語復元に関する関連研究として，入力された略語を復元規則に基づいて復元する手法が提案されている (石井直樹, 平石智宣, 延澤志保, 斎藤博昭, 中西正和 2000)．また，入力されたカタカナ語の省略形を WEB 文書を使用して推定する手法が提案されている (野呂康洋, 桝井文人, 河合敦夫 2003)．それらに対して，本手法は日本語の略語とそれに対応する元の語との対応をコーパスから取得する手法であり，入力された略語を元の語へ復元する研究ではない．そのため，コーパスに出現する各名詞が略語か原型語であるかどうかの情報は与えられていないことを手法の前提としている．また，本手法ではカタカナ語の省略形以外にも，漢字による略語と原型語との対応も取得する．しかし，3.1 節でも述べるが，表層的な文字情報による規則のみで不適切な略語候補と原型語候補との対応を排除し，略語とその原型語との対応を取得することは困難であると考える．そのため，本手法では，コーパスに含まれる名詞対から規則を適用して略語候補と原型語の候補との対応を取得し，その後，略語候補と原型語候補の類似度を統計情報を使用して計算し，類似度が高い略語候補と原型語候補との対応を略語と原型語との対応として取得する．

英語における略語と原型語との対応のコーパスからの取得の先行研究としては，一方は略語を多く含み，もう一方は略語が少ない同一ドメインの文書群を２つ用意し，略語を多く含む文書における未知語は何かの略語である可能性があるとして，もう一方の文書群から対応する名詞を抽出する手法が提案されている (Terada and Tokunaga 2001)．しかし，この手法では同一ドメインの略語を多く含むコーパスと含まないコーパスを共に用意しなければならない．そのため，そのようなコーパス対の入手が必ずしも可能であるとは限らない上，ドメインが限定されるため，そのドメインにおける略語と原型語との対応しか抽出できない．

それらの先行研究に対して，我々は，以前に単一のコーパスから略語と原型語との対応を抽出する試みを行なった (酒井浩之・増山繁 2002)．この手法は，日本語における名詞の文字情報

による規則を用いて，略語候補と原型語候補との対応をコーパスから探索する．そして，略語候補と原型語候補の名詞間類似度を計算し，ある閾値より名詞間類似度が高い対応を略語と原型語との対応として取得する．我々の既提案手法 (酒井浩之・増山繁 2002) は，単一のコーパスから名詞と略語との対応を抽出することを試みるため，ドメインに依存しない略語を抽出可能であった．また，文内の構造は用いないため，例えば括弧表現のような特殊な構造をしていなくても，略語と原型語との対応を抽出可能である．本稿で提案する手法は，我々の既提案手法を改良し，既提案手法に比べて同一の再現率において高い精度を目指したものである．

本手法と既提案手法とでは，名詞間類似度の計算手法が異なる．既提案手法は，正しい略語と原型語の対応でも低い類似度が割り当てられることがある．既提案手法では略語候補，原型語候補が出現する文に含まれる全ての名詞を取得し，それを意味属性に汎化する．そして，その意味属性に付与された重みを要素としたベクトルを生成し，ベクトル間の余弦を略語候補と原型語候補の類似度とする．しかし，略語候補や原型語候補と関連が低い名詞の意味属性をも要素とするため，多くの非零要素を持つベクトルが生成される．そのため，たとえ正しい略語と原型語の対応でも低い類似度が割り当てられることがある．それに対して，本手法は，略語候補や原型語候補に対するベクトルを生成する際に，略語候補や原型語候補を含む文書から名詞を取得する．そして，略語候補や原型語候補に関連のある名詞を抽出する処理を行い，無関係な名詞によるベクトル成分への悪影響を抑制する．評価の結果，既提案手法より同一の再現率において高い精度を達成した．また，提案した手法のより詳細な実験を行い，本手法の有効性を確認した．

2　手法の概要

本稿では，略語とそれに対応した元の語との対応関係をコーパスから自動的に獲得する手法を提案する．与えられる情報はコーパスのみであり，ある名詞が何かの名詞の略語であるといった情報は与えられない．そのため，本手法では，まず，ある名詞の略語である可能性のある名詞を取得する．以降，ある名詞の略語である可能性のある名詞を「略語候補」と定義する．そして，ある略語候補の原型語である可能性のある名詞を取得し，これを「原型語候補」と定義する．なお，各略語候補に対して，それぞれ，対応する原型語候補を獲得する．本手法は次の 2 つのステップから成る．

Step 1　名詞を構成する文字情報を使用して，コーパスから「略語候補」とその「原型語候補」との対応を獲得する．

Step 2　「略語候補」と「原型語候補」の名詞間類似度を計算し，類似度が高い略語候補と原型語候補との対応を略語とその原型語との対応とする．

Step 1 では，略語の生成規則に基づく条件を設定し，略語候補と原型語候補との対応をコーパスから取得する．しかしながら，略語の生成規則に基づく条件の適用だけでは多くの誤った略

語候補と原型語候補との対応を取得する．そのため，Step 2で略語候補と原型語候補の名詞間類似度を計算し，類似度が低い略語候補と原型語候補との対応を排除し，類似度が高い略語候補と原型語候補との対応を略語と原型語との対応として認定する．Step 1で使用する条件や規則は，我々が既提案手法において提案した条件 (酒井浩之・増山繁 2002) とほぼ同一であるが，Step 2における名詞間類似度計算手法は異なる．本稿で提案する手法は，Step 2における名詞間類似度計算手法を改良したことで，既提案手法よりも同一の再現率において高い精度を達成する．Step 1の「略語候補」と「原型語候補」の取得については3.1節で，Step 2の「略語候補」と「原型語候補」の名詞間類似度については4.1節でそれぞれ説明する．

3　略語候補とその原型語候補との対応の獲得

3.1　略語候補とその原型語候補の取得

本手法は，略語と原型語との対応を，コーパスに含まれる「略語候補」とその「原型語候補」とを比較することで獲得する．そのため，まず，「略語候補」と「原型語候補」との対応をコーパスから獲得しなければならない．本手法では，次の条件1を満たす名詞Aと名詞Bとの対応を，略語候補Aと原型語候補Bとの対応とする．

条件1　名詞Aに含まれている全ての文字が，名詞Bに同じ順序で出現する．

しかし，この条件のみでは多くの不適切な略語候補Aと原型語候補Bとの対応が取得されるため，次の2つの規則を加え，不適切な対応を排除する．

規則1　原型語候補Bには略語候補Aが連続した文字列として含まれていない．

規則2　略語候補Aと原型語候補Bの先頭の文字が同一である．

例えば，名詞A「原発」は，名詞B「原子力発電所」に対して全ての条件，および，規則を満たす．すなわち，「原発」を1文字ごとに分割すると2文字に分割される．そして「原子力発電所」は2文字全てを含み，かつ，出現順序も「原・発」の順序であり等しい (条件1)．しかも，「原子力発電所」に「原発」は連続した文字列としては含まれていない (規則1)．さらに，両名詞の先頭の文字は共に「原」であり，同一である (規則2)．「原発」と「原子力発電所」は条件，および規則を全て満たす対応であるため，「原発」と「原子力発電所」を，略語候補と原型語候補との対応として取得する．

しかし，これらの条件，および規則だけでは，略語候補が「三重」，それに対応する原型語候補が「三菱重工業」といった明らかに不適切な対応も抽出してしまう．よって，表層的な文字情報による規則のみで不適切な略語候補と原型語候補との対応を排除し，略語とその原型語との対応を取得することは困難であると考える．そこで，略語候補と原型語候補の類似度を計算し，類似度が大きい略語候補と原型語候補との対応を略語と原型語との対応として取得する．略語候補と原型語候補の類似度については4.1節で述べる．

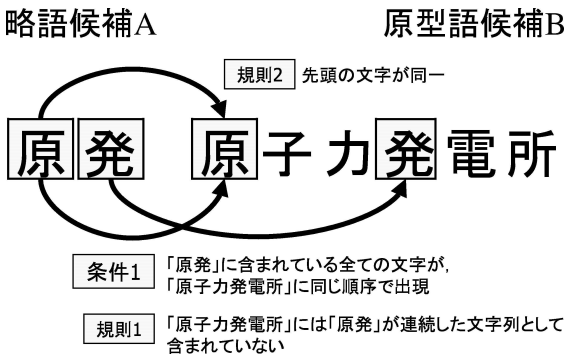

図 1　略語候補と原型語候補との対応の取得

3.2　規則 1 および規則 2 の有効性

本稿では，ある名詞から一部の文字や名詞を省略しても意味内容を保つ語を略語と定義した．よって，全ての略語と原型語との対応は条件 1 を満たす．しかしながら，全ての略語と原型語との対応は規則 1 と規則 2 を満たすわけではない．規則 1 および規則 2 は，不適切な略語と原型語との対応を排除するために導入しているが，いくつかの適切な略語と原型語との対応も同時に排除してしまう．例えば,「スト」と「ストライキ」は適切な略語と原型語との対応であるが規則 1 を満たさない．また,「空母」と「航空母艦」は適切な略語と原型語との対応であるが規則 1 と規則 2 を共に満たさない．本手法では不適切な略語候補と原型語候補との対応を排除するために規則 1 と規則 2 を導入しているが，それらが有効であるかどうかを予備実験を行って調査した．予備実験の方法を以下に示す.

Step 1:　コーパスから条件 1 を満たす略語候補と原型語候補との対応を全て取得する.

Step 2:　取得した対応を，次の 4 つの集合，$C_{00}, C_{01}, C_{10}, C_{11}$ に分類する.

C_{00}:　規則 1 と規則 2 を共に満たさない略語候補と原型語候補との対応の集合．(すなわち，原型語候補 B は略語候補 A を連続した文字列として含み，かつ，略語候補 A と原型語候補 B の先頭の文字が同一ではないとき，かつ，そのときのみ，A と B との対応は C_{00} に属する.「空母」と「航空母艦」などが該当する．)

C_{01}:　規則 1 を満たさないが規則 2 を満たす略語候補と原型語候補との対応の集合．(すなわち，原型語候補 B は略語候補 A を連続した文字列として含み，かつ，略語候補 A と原型語候補 B の先頭の文字が同一であるとき，かつ，そのときのみ，A と B との対応は C_{01} に属する.「スト」と「ストライキ」などが該当する．)

C_{10}:　規則 1 を満たすが規則 2 を満たさない略語候補と原型語候補との対応の集合．(すなわち，原型語候補 B は略語候補 A を連続した文字列として含まず，かつ，略語候補 A と原型語候補 B の先頭の文字が同一ではないとき，かつ，そのときのみ，

表 1　規則1，および，規則2による判定数

集合	規則1	規則2	取得数	判定数
C_{00}	0	0	968302	2
C_{10}	1	0	524396	1
C_{01}	0	1	545276	8
C_{11}	1	1	372576	16

取得数: C_{xx} に属する対応の数.
判定数: 判定された適切な対応の数

A と B との対応は C_{10} に属する.「安保理」と「国連安全保障理事会」などが該当する．)

C_{11}:　規則1と規則2を共に満たす略語候補と原型語候補との対応の集合．(すなわち，本手法で取得される対応の集合.「原発」と「原子力発電所」などが該当する．)

Step 3:　各集合から，100個の対応を無作為に抽出し，人手で適切な略語と原型語との対応を判別する．

Step 4:　抽出した100個に含まれる適切な略語と原型語との対応の数を調査する．

略語と原型語との対応関係を取得する対象のコーパスとして，93年の日経新聞記事1月1日から6月30日までの約84,905記事を採用した．その結果，2,410,550個の条件1のみを満たす略語候補と原型語候補との対応が取得された．表1に予備実験の結果を示す．なお，表1では，規則を満たす場合を1，満たさない場合を0で表記した．

3.3　予備実験の考察

表1から，C_{11} が最も多くの適切な略語と原型語との対応を含んでいることが分かる．そのため，規則1および規則2は不適切な略語候補と原型語候補との対応を排除するのに有効であると考える．

C_{01} は,「スト」と「ストライキ」のような適切な略語と原型語との対応を含むが,「アップ」と「アップル」のような不適切な対応も多く含む．同様に，C_{00} は,「空母」と「航空母艦」のような適切な略語と原型語との対応を含むが,「ＥＣ」と「ＮＥＣ」のような不適切な対応も多く含む．本手法では，略語候補と原型語候補の名詞間類似度を計算し，高い類似度をもつ略語候補と原型語候補との対応を略語と原型語との対応と認定する．しかしながら，類似度計算を行う前に不適切な略語候補と原型語候補がなるべく排除されている方が高い精度が期待できるため，本手法では規則1と規則2を導入した．

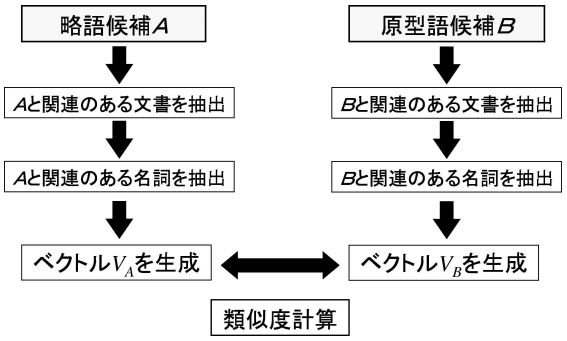

図 2　略語候補と原型語候補の類似度計算手法の概要

4　略語と原型語との対応関係の認定

4.1　略語候補と原型語候補の類似度

取得した略語候補と原型語候補との対応から略語と原型語との対応を認定するため，略語候補と原型語候補の類似度をベクトル空間法 (Salton 1988)(Baeza-Yates and Ribeiro-Neto 1999) に基づく手法で計算する．本節では，略語候補と原型語候補の類似度計算手法について述べる．まず，コーパス中の略語候補を含んでいる文書に重みを付与して順位付けを行ない，その上位文書を抽出することで略語候補と関連のある文書集合を抽出する．次に，その文書集合に含まれている名詞に対して重みを付与して順位付けを行ない，その上位の名詞を抽出することで略語候補と関連のある名詞を抽出する．そして，その名詞の重みを要素としたベクトルを生成する．それにより，略語候補と関連のある名詞に付与された重みを要素としたベクトルが生成される．(なお，文書および名詞への重み付与手法は後述する．) 同様の処理を行ない，原型語候補と関連のある名詞を抽出し，その名詞の重みを要素としたベクトルを生成する．そして，2つのベクトルの余弦を計算し，余弦がある閾値以上の略語候補と原型語候補との対応を略語と原型語との対応と判定する．以下に手法を示す．

Step 1　略語候補 A を含む文書集合 D_A をコーパスから抽出する．

Step 2　文書集合 D_A の要素である各文書 $d \in D_A$ の重み $Wd(d, A)$ を，以下の式1で計算する．

$$Wd(d, A) = \frac{tf(A, d)}{1 + \log(atf(d))} \cdot \frac{1 + nl(d) - nlf(A, d)}{nl(d)} \tag{1}$$

ただし，

$tf(A, d)$: 文書 $d \in D_A$ における略語候補 A の出現頻度，

$atf(d)$: 文書 $d \in D_A$ に含まれる全ての名詞の総出現頻度，

$nl(d)$: 文書 $d \in D_A$ を構成する文の数，

$nlf(A,d)$: 文書 $d \in D_A$ において，略語候補 A が最初に出現する文番号，(もし，略語候補 A が文書 d の第一文に出現するならば，$nlf(A,d) = 1$．)

Step 3　Step 2によって付与された重みが大きい上位 n 文書を抽出し，その文書集合を S_A とする．抽出された n 個の文書を略語候補 A の関連文書とする．

Step 4　文書集合 S_A に含まれる各名詞 t_j に対して，式2によって計算される重み $Wt(t_j, S_A)$ を付与する．

$$Wt(t_j,S_A) = \frac{TF(t_j,S_A)}{1+\log(ATF(S_A))} \cdot \log\frac{|\Delta|}{df(t_j,\Delta)} \times max_{d\in S_A}\frac{1+nl(d)-nlf(t_j,d)}{nl(d)} \quad (2)$$

ただし，

$TF(t_j,S_A)$: 文書集合 S_A における名詞 t_j の出現頻度．式3のように計算される．

$$TF(t_j,S_A) = \sum_{d\in S_A} tf(t_j,d) \quad (3)$$

$tf(t_j,d)$ は，文書集合 S_A の要素である文書 d における名詞 t_j の出現頻度である．

$ATF(S_A)$: 文書集合 S_A に含まれる全ての名詞の総出現頻度．

Δ: 全体の文書集合．すなわち，略語と原型語との対応関係を取得するために使用するコーパス．

$df(t_j,\Delta)$: 全体の文書集合 Δ における名詞 t_j の文書頻度．

Step 5　Step 4によって付与された重み $Wt(t_j,S_A)$ が大きい上位 m 個の名詞を抽出する．抽出された m 個の名詞を略語候補 A の関連名詞とする．

Step 6　略語候補 A における，Step 5で抽出された各関連名詞 t_j に対応する要素を，それぞれ，その重み $Wt(t_j,S_A)$ とし，それ以外の要素を0としたベクトル $\boldsymbol{V}_A$ を生成する．

Step 7　原型語候補 B についてもStep 1からStep 6の処理を行ない，ベクトル $\boldsymbol{V}_B$ を生成する．

Step 8　ベクトル $\boldsymbol{V}_A$ とベクトル $\boldsymbol{V}_B$ の余弦を計算し，名詞間類似度 $sim(A,B)$ とする．

$$sim(A,B) = \frac{\boldsymbol{V}_A \cdot \boldsymbol{V}_B}{|\boldsymbol{V}_A||\boldsymbol{V}_B|} \quad (4)$$

Step 9　名詞間類似度 $sim(A,B)$ がある閾値以上の略語候補 A と原型語候補 B との対応を略語と原型語との対応と判定する．

本手法は，略語候補 A と原型語候補 B が同一の意味を持つのであれば同一の内容の文書で使われることが多く，略語候補，原型語候補に関連している名詞には共通した名詞が多いという仮定に基づく．

式1の1番目の項の分子は，略語候補 A が文書 d に多く出現している場合に大きな重みを与えるための項である．これは，略語候補 A が多く出現している文書は略語候補 A と関連のある

文書である可能性があるという仮定に基づく．しかし，文書dが長ければ多くの語が含まれるため，必然的に文書dに略語候補Aを多く含む可能性が高くなる．そして，長い文書であれば多くの内容が含まれているため，略語候補Aを多く含んでいたとしても，その文書は略語候補Aとの関連が低い可能性がある．式1の1番目の項の分母は，文書の長さによる影響を軽減するために導入した項である．長い文書であれば文書dに含まれる全ての名詞の総出現頻度$atf(d)$は高くなるため，長い文書における重みを文書の長さに応じて小さくできる．式1の2番目の項は，略語候補Aが文書dの最初に出現している文書に対して大きな重みを与えるための項である．これは，文書中の最初の文は重要な文である可能性が高く，そのような文に略語候補Aが含まれていれば，その文書は略語候補Aと関連のある文書である可能性があるという仮定に基づく．

式2の1番目と2番目の項は，文書集合S_Aを1つの文書とした場合における$tf \cdot idf$法(Salton 1988)(Baeza-Yates and Ribeiro-Neto 1999)に基づく．式2の1番目の項の分子は，文書集合S_Aに多く出現している名詞に対して大きな重みを与えるための項である．これは，文書集合S_Aに多く出現している名詞は，その文書集合と関連のある名詞である可能性があるという仮定に基づく．分母は，式1の1番目の項の分母と同じく，文書の長さによる影響を軽減するために導入した項である．しかし，名詞の出現頻度のみでは，多くの文書に出現する一般的な名詞に対して高い重みを与えてしまうため，それを補正するために式2の2番目の項を導入する．もし名詞t_jが，多くの文書に出現するような一般的な語であった場合，$df(t_j, \Delta)$が大きくなるため，そのような名詞の重みは低くなる．式2の3番目の項は，名詞t_jが文書の最初に出現しているならば大きな重みを与えるための項である．これは，文書中の最初の文は重要な文である可能性が高く，そのような文に含まれている名詞も同じくその文書にとって重要な名詞である可能性があるという仮定に基づく．

4.2　省略名詞による制限

ここで，精度を向上させるために省略名詞について各種の制限を加える．省略名詞とは，原型語候補に含まれる普通名詞の中で略語候補を構成する文字を含まない名詞と定義する．例えば，略語候補「バブル崩壊」と原型語候補「バブル経済崩壊」では「経済」が省略名詞となる．なぜなら,「バブル経済崩壊」には,「バブル」「経済」「崩壊」の3つの普通名詞が含まれるが,「経済」を構成する文字は「バブル崩壊」には含まれていないからである．なお，本手法の実装にあたり，形態素解析器としてJUMAN [1] version 3.5を採用したが，複合名詞の分割はJUMAN version 3.5に従っている．

本手法では，この省略名詞に対して，次の2つの制限を加える．

制限1　原型語候補の中に省略名詞が複数存在していれば，その略語候補Aと原型語候補Bとの対応を略語と原型語との対応と判定しない．($sim(A, B) = 0$とする．)

1 http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

制限2　原型語候補である複合名詞の最後の位置にある普通名詞(例えば,「バブル経済崩壊」なら「崩壊」)が省略名詞であるならば，その略語候補Aと原型語候補Bとの対応を略語と原型語との対応と判定しない．($sim(A,B)=0$とする．)

制限1について説明する．省略名詞が存在している原型語候補が対応する略語候補と意味が同一であると判断できるのならば，その省略名詞を推定できる必要がある．しかし，省略名詞が複数存在しているならば，それらを全て推定できる可能性は低い．例えば，本手法では「民間会社」と「民間生命保険会社」が略語候補と原型語候補との対応として取得される．この場合,「生命」と「保険」が省略名詞となる．しかし,「民間会社」から「生命」と「保険」を推定できず，よって，このような略語と原型語との対応が生じる可能性は低い．そのため，制限1を加える．

制限2について説明する．原型語候補の最後の普通名詞が省略されると意味が異なる名詞になることがある．例えば,「売上」と「売り上げ不振」では,「不振」が省略名詞であるが,「不振」が省略されることで意味が異なる名詞になる．そのため，制限2を加える．

4.3　予備実験

本手法では，略語候補，原型語候補に対するベクトルを生成するために，Step 3においてn個の関連文書，Step 5においてm個の関連名詞を取得する必要がある．このパラメータn，mの値を決定するために予備実験を行う．まず，略語と原型語との対応関係を取得するコーパスとして，93年の日経新聞記事1月1日から6月30日までの約$84,905$記事を採用し，このコーパスから無作為に500個の略語候補と原型語候補との対応を取得した．なお，コーパスから取得した略語候補と原型語候補との対応は，3.2節で示したC_{11}に属する略語候補と原型語候補との対応に相当する．そして，取得した対応を人手によって判定した結果を正解データとして，精度P，再現率Rを計算する．

なお，再現率を測定するためには，コーパス中に存在する全ての略語と原型語との対応を人手で判定して正解データを作成する必要がある．しかし，それは困難な作業である．なぜなら，表層的な文字情報による幾つかの規則を課しても$372,576$個もの略語候補と原型語候補との対応が取得されてしまうため，コーパス中に存在する全ての略語と原型語との対応を人手で判定することは不可能であるからである．そのため，本稿では，正解データの再現性を測る尺度という意味で再現率を用いる．

精度P，再現率Rは以下のように定義される．

$$R = C/X\ (\times 100\%), \qquad P = C/Y\ (\times 100\%),$$

ただし，

C:　本手法によって判定された略語と原型語との対応と正解データにおける略語と原型語との対応で，一致する略語と原型語との対応の数

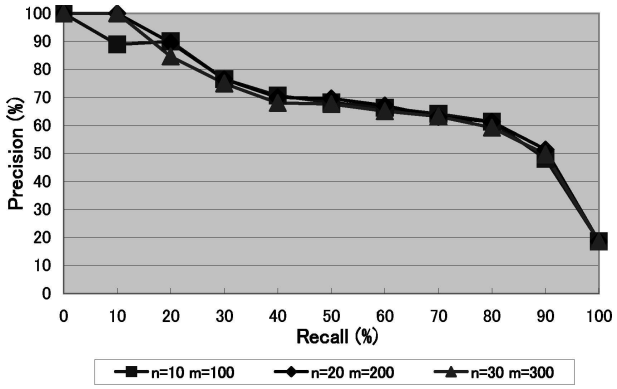

図 3　パラメータを変化した場合の再現率ー精度グラフ

X:　正解データにおける略語と原型語との対応の数

Y:　本手法によって判定された略語と原型語との対応の数

パラメータn，および，mの値を決定するために，$n = 10, m = 100$の場合，$n = 20, m = 200$の場合，$n = 30, m = 300$の場合に関して精度，再現率を測定した．図3に結果を示す．本手法は，閾値によって精度，および，再現率が変化するので，再現率ー精度グラフで結果を示す．実験結果から，パラメータn，および，mによって結果が大きく変化しないことが分かる．よって，本手法では$n = 20$，$m = 200$とする．

5　手法の実装

本手法を実装して略語と原型語との対応関係の獲得を行なった．コーパスとして93年の日経新聞記事1月1日から6月30日までの約84,905記事を採用し，形態素解析器としてJUMAN[2] version 3.5を採用した．表2に，閾値を0.3に設定した場合に本手法によって取得した略語と原型語との対応をいくつか示す．

6　評価

6.1　評価実験および結果

本手法を精度，再現率で評価する．評価方法を以下に示す．

Step 1　1,000個の略語候補と原型語候補との対応を無作為に取得する．

Step 2　取得した略語候補と原型語候補との対応の中から，適切な略語と原型語との対応を人手で判定し，正解データを作成する．

2 http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

表 2　本手法によって取得した略語と原型語との対応

略語	原型語	類似度
原発	原子力発電所	0.39
生保	生命保険	0.35
東電	東京電力	0.37
公取委	公正取引委員会	0.58
安保理	安全保障理事会	0.47
スパコン	スーパーコンピューター	0.44
団交	団体交渉	0.68
特許料	特許使用料	0.59
バブル崩壊	バブル経済崩壊	0.34
関空	関西国際空港	0.58

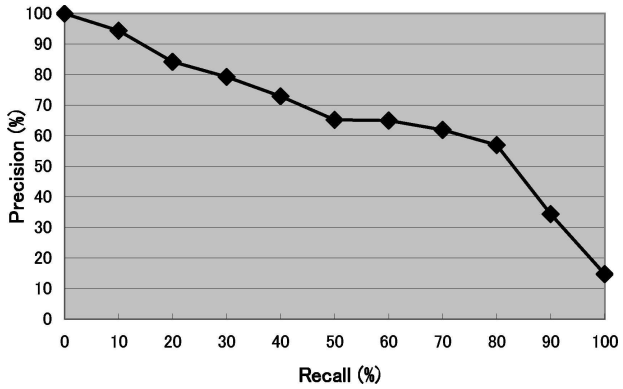

図 4　本手法の再現率ー精度グラフ

Step 3　正解データから精度，再現率を計算する．

なお，Step 1における略語候補と原型語候補との対応は，3.2節で示したC_{11}に属する略語候補と原型語候補との対応に相当し，採用したコーパスからの取得数は372,576個である．その中から1,000個の対応を無作為に抽出し正解データを作成したが，5人の工学系の大学生および大学院生に略語候補と原型語候補との対応を判定してもらい，3人以上が適切であると判定した対応を正解とした．その結果，正解データには147個の適切な略語と原型語との対応が存在した．図4に結果を示す．また，表3に，閾値を0.1から0.8まで変化させた場合の精度，再現率を示す．また，採用したコーパスから取得された対応の数も並記する．

表 3　本手法の精度，再現率

閾値	精度 (%)	再現率 (%)	取得数
0.1	57.2	75.5	37318
0.2	64.9	60.5	21528
0.3	69.1	44.2	14387
0.4	73.4	31.9	10120
0.5	79.1	23.1	7208
0.6	84.8	19.0	5164
0.7	94.1	10.9	3766
0.8	100	6.1	2745

6.2　単語頻度のみの手法との比較

本手法と単語頻度 (tf) のみを使用した手法とを比較する．単語頻度 (tf) のみを使用した手法では，式1の $Wd(d, A)$，および式2の $Wt(t_j, S_A)$ が以下のように変更される．

$$Wd(d, A) = tf(A, d),$$

$$Wt(t_j, S_A) = \sum_{d \in S_A} tf(t_j, d),$$

本手法における文書や文書に含まれる語の重みは $tf \cdot idf$ 法 (Salton 1988)(Baeza-Yates and Ribeiro-Neto 1999) に基づく手法で付与している．しかし，Terada らは文献 (Terada and Tokunaga 2001) で tf のみを使用したベクトルで英語の略語と原型語との対応をコーパスから取得している．(文献 (Terada and Tokunaga 2001) と本手法との違いは1章で述べたが，文献 (Terada and Tokunaga 2001) は，一方は略語を多く含み，もう一方は略語が少ない同一ドメインのコーパスを２つ用意し，それらのコーパスから略語と原型語との対応を取得する．それに対して，本手法はドメインが限定されていない単一コーパスから略語と原型語との対応を取得する．) そこで，本手法と tf のみを使用した手法とを比較する．結果を図5に示す．

6.3　文字省略型と名詞省略型の比較

本稿では，ある名詞から一部の文字や名詞を省略しても意味内容を保つ語を略語と定義した．例えば,「生保」 は原型語「生命保険」から一部の文字が省略されることで生成される．それに対して,「バブル崩壊」は原型語「バブル経済崩壊」から名詞「経済」が省略されることで生成される．本節では，原型語から一部の文字が省略されることで生成される略語を「文字省略型」と定義し，原型語から名詞が省略されることで生成される略語を「名詞省略型」と定義する．そして,「文字省略型」と「名詞省略型」の略語と原型語との対応を比較する．結果を図6に示す．また，表4に，閾値を 0.1 から 0.8 まで変化させた場合の精度，再現率，取得数をそれぞれ示す．なお，6.1節で作成した正解データには適切な「文字省略型」の対応が 77 個，適切な

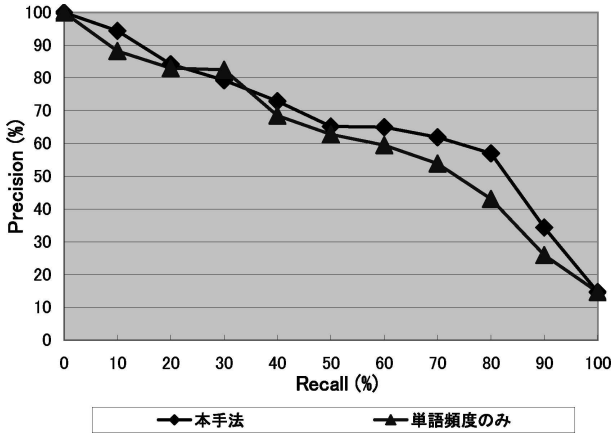

図 5　単語頻度 (tf) のみを使用した手法との比較

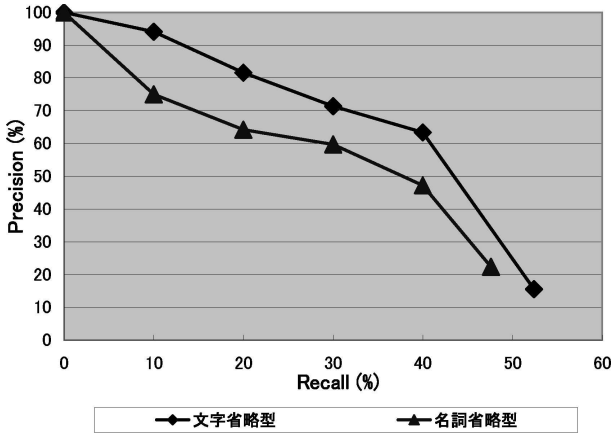

図 6　文字省略型と名詞省略型の比較

「名詞省略型」の対応が70個含まれており，「文字省略型」の再現率の上限は52.4%，「名詞省略型」の再現率の上限は47.6%となる．

6.4　$C_{00} \cup C_{01} \cup C_{10} \cup C_{11}$ との比較

3.2節では，条件1を満たす略語候補と原型語候補との対応を，規則1および規則2を満たす場合，満たさない場合で4つの集合 C_{00}，C_{01}，C_{10}，C_{11} に分類した．本手法では，C_{11} に属する略語候補と原型語候補との対応から略語と原型語との対応を取得したが，C_{00}，C_{01}，C_{10} にも適切な略語と原型語との対応がいくつか属している．そこで，規則1と規則2の有効性を精

表 4　文字省略型と名詞省略型の比較

	文字省略型			名詞省略型		
閾値	精度 (%)	再現率 (%)	取得数	精度 (%)	再現率 (%)	取得数
0.1	58.5	42.2	25062	55.7	33.3	12256
0.2	67.1	37.4	14695	61.8	23.1	6833
0.3	74.1	29.3	9884	61.1	15.0	4503
0.4	75.6	21.1	7020	69.6	10.9	3100
0.5	84.4	18.4	5062	63.6	4.8	2146
0.6	84.0	14.3	3718	87.5	4.8	1446
0.7	92.9	8.84	2771	100	2.0	995
0.8	100	5.4	2077	100	0.7	668

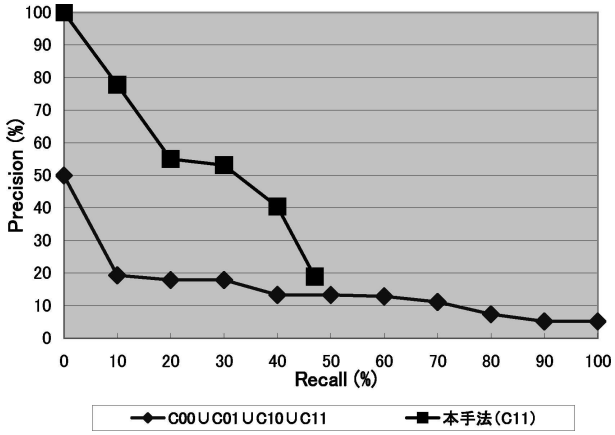

図 7　$C_{00} \cup C_{01} \cup C_{10} \cup C_{11}$ との比較

度，再現率で評価するために，条件1のみを満たす略語候補と原型語候補との対応の集合 (すなわち，$C_{00} \cup C_{01} \cup C_{10} \cup C_{11}$) と，本手法で対象としている集合 (すなわち，$C_{11}$) から，それぞれ略語と原型語との対応を取得し，比較評価する．評価のために，条件1のみを満たす略語候補と原型語候補との対応から無作為に1000個の対応を抽出し，人手で判定を行い正解データを作成する．そして，正解データから精度，再現率を計算する．結果を図7に示す．作成した正解データには52個の適切な略語と原型語との対応が含まれていた．その中で，C_{11} に属する対応は25個であった．そのため，本実験では，条件1のみを満たす略語候補と原型語候補との対応 ($C_{00} \cup C_{01} \cup C_{10} \cup C_{11}$) における，本手法 ($C_{11}$) による正解データの再現率の上限は48% となる．

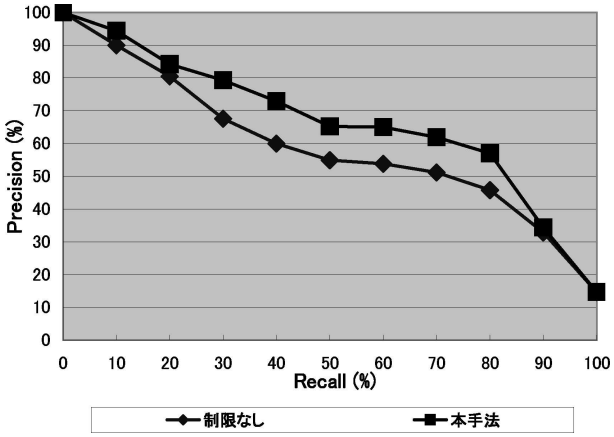

図 8　省略名詞による制限を課していない場合との比較

6.5　省略名詞による制限の効果

本手法では，名詞省略型の略語候補と原型語候補との対応に対して，4.2節で説明した省略名詞による制限を課している．この制限の有効性を評価するため，4.2節で示した制限1および制限2を課していない場合と本手法とを比較評価する．結果を図8に示す．

6.6　コーパス量の変更による本手法の影響

本手法はコーパスから略語と原型語との対応を自動的に獲得するための手法である．そのためには，ある程度のコーパス量が必要になるが，コーパス量を大きくしても本稿で設定したパラメータ $n = 20$，$m = 200$ のままでも精度，再現率に大きな影響がないか実験を行った．実験では，コーパス量を半年，1年，1年半と変更した場合における再現率，精度を比較した．なお，コーパス量半年は93年の1月1日から6月30日までの日経新聞記事，コーパス量1年は92年の7月1日から93年の6月30日までの日経新聞記事，コーパス量1年半は92年の1月1日から93年の6月30日までの日経新聞記事である．実験結果を図9に示す．

7　既提案手法との比較

1節で述べたように，本稿で提案する手法は，我々による既提案手法(酒井浩之・増山繁 2002)(以降，既提案手法と表記) と同一の再現率において高い精度を達成することを目指したものである．本手法と既提案手法(酒井浩之・増山繁 2002) とは，3.1節で説明した略語候補と原型語候補を取得するための表層的な文字情報による条件や規則はほぼ同一である．(ただし，4.2節で説明した省略名詞による制限は課していない．) しかし，略語候補と原型語候補から略語

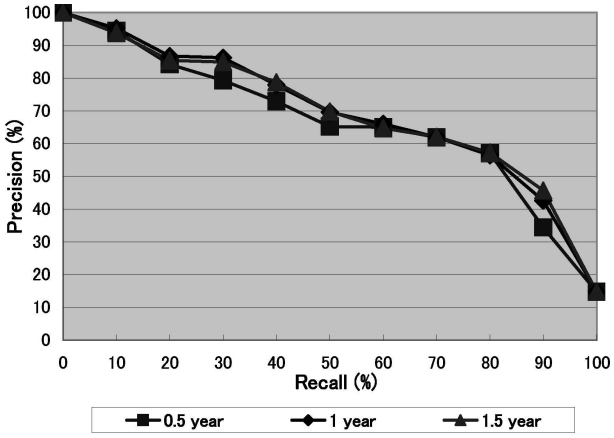

図 9　コーパス量の変更による影響

と原型語との対応を取得するための名詞間類似度の計算手法が，本手法では既提案手法と異なる．既提案手法も略語候補と原型語候補のベクトルをそれぞれ生成し，2つのベクトルの余弦を計算しているが，ベクトルを構成する要素が異なる．本手法は，略語候補や原型語候補が出現する文書に含まれる名詞を取得し，その重みを要素としてベクトルを生成する．その際，略語候補や原型語候補と関連のある名詞を抽出し，関連のない名詞を排除する処理を行なう．もし，略語候補や原型語候補が出現する文書に含まれる全ての名詞の重みを要素とすれば，多くの非零要素を含むベクトルが生成される．その場合，そのベクトルは略語候補や原型語候補の特徴を表すことができず，たとえ，略語候補と原型語候補が正しい略語と原型語との対応であったとしても低い類似度が割り当てられてしまう．そこで，本手法では，ベクトル生成の際に略語候補や原型語候補に関連のある名詞に絞り込み，関連のない名詞を排除することで，多くの非零要素を含むベクトルの生成を防ぐ．

それに対して，既提案手法では略語候補，原型語候補が出現する文に含まれる全ての名詞を取得し，それを意味属性に汎化する．そして，その意味属性に付与された重みを要素としたベクトルを生成する．もし，複数の意味属性が割り当てられる名詞である場合は，割り当てられる全ての意味属性を採用する．(なお，意味属性辞書として「日本語語彙大系」(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) 1997) を使用している．) 既提案手法では，略語候補，原型語候補が出現する文に含まれる全ての名詞を取得するため，略語候補や原型語候補と関連のない名詞を排除する処理を行なっていない．なぜなら，略語候補や原型語候補が出現する文に含まれる全ての名詞を取得したとしても，本手法のように略語候補や原型語候補が出現する文書に含まれる名詞を取得した場合に比べて，取得できる名詞があまりに少ないからである．つまり，略語候補や原型語候補と関連のない名詞を排除できるほど

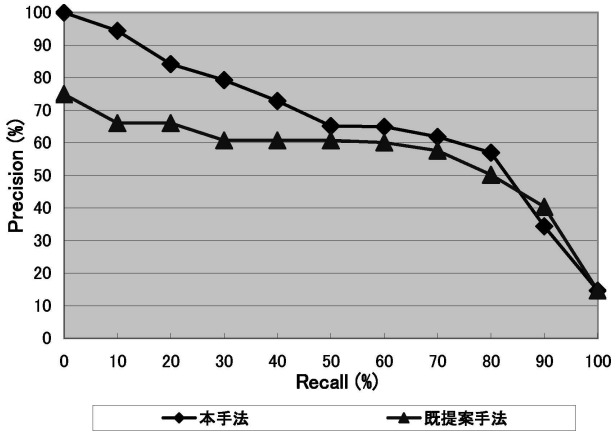

図 10　本手法と既提案手法との比較実験結果

の統計情報が得られない．そこで，既提案手法では，略語候補，原型語候補が出現する文に含まれる全ての名詞を意味属性で汎化し，その意味属性に付与された重みを要素としたベクトルを生成した．また，本手法では，意味属性辞書が不要である点も既提案手法に比べて利点となる．実験結果を図10に示す．

8　考察

8.1　評価結果

図5から，本手法は単語頻度のみの手法よりも同一再現率において高い精度を達成した．本手法は，略語候補や原型語候補に関連のある名詞を抽出してベクトルの要素としている．ここで，略語候補や原型語候補に関連のある名詞としては，多くの文書に出現するような一般的な名詞ではない方が望ましい．なぜなら，そのような名詞がベクトルの非零要素となった場合，適切ではない略語候補と原型語候補との対応であっても類似度が高くなる原因になるからである．しかし，単語頻度のみでは一般的な名詞が略語候補や原型語候補に関連のある名詞として抽出され，各ベクトルの非零要素になる可能性がある．そのため，本手法の方が良い結果であったと考える．

図6から，本手法では名詞省略型の略語と原型語との対応よりも文字省略型の略語と原型語との対応の方が，同一再現率において高い精度を達成した．また，表4から，文字省略型の略語と原型語との対応の方が同一閾値において取得数が多いことが分かる．従って，より高い精度が必要であるならば，文字省略型の略語と原型語との対応のみを取得するように制限を加えることで対応できると考える．

図7から，$C_{00} \cup C_{01} \cup C_{10} \cup C_{11}$ に属する略語と原型語との対応(条件1のみを満たす略語と原型語との対応)よりも C_{11}(本手法)に属する略語と原型語との対応の方が，同一再現率において高い精度を達成した．条件1のみを満たす略語候補と原型語候補との対応 ($C_{00} \cup C_{01} \cup C_{10} \cup C_{11}$) は多くの不適切な対応を含み，略語候補と原型語候補の類似度が低い対応を排除しても高い精度が達成できなかった．しかし，限定されたドメインの文書集合に対して本手法を適用するならば，条件1のみを満たす略語候補と原型語候補との対応から略語と原型語との対応を取得しても，ある程度の高い精度が達成できると考える．規則1および規則2は，不適切な略語候補と原型語候補との対応を排除するために導入しているが，限定されたドメインの文書集合から略語候補と原型語候補との対応を取得すれば，取得される不適切な対応の数が減少すると考える．例えば，略語候補「パン」と原型語候補「パソコン」との対応は不適切であるが，コンピュータ関連に限定された文書集合からこのような対応は取得されないと考える．そして，不適切な略語候補と原型語候補との対応が取得されなければ，規則1および規則2を導入しなくても高い精度が達成できると考える．

図8から，本手法は制限1および制限2を課していない手法に比べて同一再現率で高い精度を達成した．よって，制限1および制限2を課すことは有効であると考える．例えば，制限1および制限2を課していなければ,「米国経済」と「米国際経済研究所」が略語と原型語との対応として取得された．これは明らかに不適切な対応であるが，制限2を課すことでこのような不適切な対応を排除することができた．

図9から，コーパス量が極端に少なくなければ，コーパス量の変更によって精度，再現率に大きな変化がないことが分かる．本手法は，略語候補，および，原型語候補に関連のある名詞をコーパスから抽出し，その名詞に付与された重みを要素としたベクトルを生成する．そして，そのベクトル間の余弦がある閾値以上の略語候補と原型語候補との対応を，略語と原型語との対応と判定する．しかし，コーパス量が変更されても本手法で抽出される略語候補，および，原型語候補に関連のある名詞の上位は大きな変化がなく，そのため，精度，再現率も大きな変化がないと考える．

8.2　既提案手法との比較

図10から，本手法は我々の既提案手法(以降，既提案手法とする．)より優れた結果が得られることが分かる．本手法は再現率が40%まで精度70%を維持している．しかし，既提案手法は，再現率40%において精度が60%まで低下している．その理由として以下が考えられる．

既提案手法は，略語候補や原型語候補が出現する文に含まれている全ての名詞の意味属性に対して重みを計算し，それらを要素としたベクトルを生成する．文に含まれている名詞だけでは，略語候補のベクトルと原型語候補のベクトルに共通する要素数が少ないため，文に含まれている名詞を全て取得し，その名詞を意味属性に汎化する必要があった．そのため，略語候補

や原型語候補と関連が低い名詞の意味属性をも要素とし，多くの非零要素を持つベクトルが生成される．そのような名詞の意味属性は，略語候補と原型語候補とで共通した要素となる可能性は低いが，もし，略語候補と原型語候補との対応が正解であった場合に，略語候補と原型語候補の名詞間類似度が小さくなる原因になる．そのため，既提案手法は，正しい略語と原型語との対応でも低い名詞間類似度が割り当てられることがある．したがって，再現率を上げるためには低い閾値を設定する必要があり，それにともなって，多くの誤判定が生じる．例えば，本手法は「省エネ対策」と「省エネルギー対策」を略語と原型語との対応として精度80%以上を達成できる閾値で取得できた．しかしながら，既提案手法では精度60%まで閾値を低く設定しないと取得できなかった.「省エネ対策」や「省エネルギー対策」は，環境問題に関する文書や国の政策に関する文書など，様々な内容の文書で使用される語である．したがって，既提案手法では,「省エネ対策」や「省エネルギー対策」のベクトルには多くの非零要素が含まれ，類似度が低下したと考える．

それに対して，本手法は略語候補や原型語候補を含む文書を順位付けし，順位が低い文書を排除することで略語候補や原型語候補と関連のある文書を抽出する．さらに，その文書に含まれている名詞を順位付けすることによって，略語候補や原型語候補と関連が高い名詞を抽出し，その名詞に付与された重みを要素とするベクトルを生成する．こうして，本手法は，略語候補や原型語候補に対するベクトルを生成する際に略語候補や原型語候補に関連のある名詞を抽出する処理を行い，無関係な名詞によるベクトル成分への悪影響を抑制することができた．そのため，既提案手法では取得できなかった略語と原型語との対応が取得でき，同一の再現率において既提案手法よりも高い精度を達成できたと考える．

しかし，本手法と既提案手法とでは，再現率50%以上では精度の向上が5%程度しかない．その理由を以下のように考える．本手法も既提案手法も，略語候補と原型語候補との対応から略語と原型語との対応を獲得するために，コーパスからの統計情報を必要とする．そのためには，略語候補，原型語候補，共に，コーパス中にある程度存在している必要がある．しかしながら，正解データには低頻度の略語と低頻度の原型語との対応も含まれ，そのような対応は本手法でも取得することが困難である．そのため，本手法と既提案手法では，再現率50%以上においての精度の向上が少なくなったと考える．

再現率が高ければ，多くの略語と原型語との対応を獲得できる．しかしながら，コーパスから自動的に獲得した略語と原型語との対応の精度が低ければ人手による修正が必要となり、それには大きな労力がかかる．特に，ある原型語に対する略語は次々と新しく発生するため，人手による修正の労力は無視できない．例えば，今回の実験で使用したコーパスは93年の日経新聞記事だが，当時まだ開港していない「関西国際空港」の略語「関空」が既に発生していることが，本手法によって獲得された略語と原型語との対応を示した表2から分かる．また,「バブル崩壊」と「バブル経済崩壊」の例も93年の時点では比較的新しい略語だと考える．このように，

表 5　本手法による誤判定

略語	原型語	類似度	誤判定の分類
送信	送受信	0.33	1
ドル	ドイツマルク	0.45	2
大権	大統領権限	0.36	2
事情	事前情報	0.56	2
私学	私立中学	0.31	1
予算案	予算修正案	0.68	1
加工業種	加工組み立て業種	0.60	1

略語は次々と新しく発生するため人手による修正の労力は無視できず，再現率だけでなく精度も重要になると考える．図10によると，既提案手法は70%以上の精度を達成するためには再現率が10%以下である必要があることが分かる．しかし，本手法では再現率40%まで精度70%以上を保っており，本手法は既提案手法より有効であると考える．

8.3　本手法による誤り分析

本手法の誤り分析を行う．本手法による誤判定は，次の2つに分類できる．

誤判定1:　原型語候補が略語候補の意味を内包している場合，もしくは，略語候補が原型語候補の意味を内包している場合．

誤判定2:　原型語候補，略語候補の意味が異なる場合．

例えば，本手法は「送信」と「送受信」を略語と原型語との対応と認定した.「送受信」には「送信」の意味も内包しているため，この誤判定は誤判定1に属する．一方，本手法は「ドル」と「ドイツマルク」を略語と原型語との対応と認定した.「ドル」と「ドイツマルク」は，3.1節で説明した条件や規則を全て満たし通貨の名称という共通点があるが,「ドル」は「ドイツマルク」の略語ではなく意味が異なる．そのため，この誤判定は誤判定2に属する．表5に，閾値を0.3に設定した場合の本手法による誤り判定の例を幾つか示す．表6に，閾値を0.3に設定した場合における本手法の誤判定を，誤判定1および誤判定2に分類した結果を示す．なお，表6には，誤判定された略語と原型語との対応を6.3節で説明した「文字省略型」「名詞省略型」の2つに分類し，それぞれの誤判定数を示す．この場合における精度は69.1%，再現率は44.2%であった．表6から,「名詞省略型」の誤判定1が多いことが分かる．例えば,「加工業種」と「加工組み立て業種」との対応は,「名詞省略型」の誤判定1に属する．本手法では，略語候補と原型語候補の名詞間類似度を計算し，類似度が高い略語候補と原型語候補を略語と原型語との対応として取得する．その際，略語候補と原型語候補のそれぞれに対して関連のある名詞に割り当てられる重みを要素としたベクトルを生成し，そのベクトル間の余弦を計算することで，略語候補と原型語候補の類似度を計算する．しかし，原型語候補が略語候補の意味を内包している場合，多

表 6　本手法による誤判定数

	文字省略型	名詞省略型	総数
誤判定 1	3	13	16
誤判定 2	9	4	13
総数	12	17	29

くの共通した関連名詞が取得されるため，高い名詞間類似度が割り当てられてしまう．よって，誤判定1に該当する略語候補と原型語候補との対応を，本手法の名詞間類似度計算法によって判定することは困難であると考える．しかしながら，このような誤判定を表層的な文字情報による制限で排除することも問題があると考える．例えば，原型語候補を普通名詞に分割し，その先頭文字のみで構成する略語候補のみを対象にするといった制限を課すことが考えられるが，そのような制限を増やしていけば取得数が減少すると考える．また，そのような制限を課したとしても「事情」と「事前情報」との対応のような誤判定は排除できない．これらの誤判定への対応は今後の課題とする．

9　むすび

本稿では略語とその原型語との対応関係をコーパスから自動的に獲得する手法を提案した．本手法は，略語候補とそれに対応している原型語候補を，それらを構成している文字情報から取得する．そして，略語候補と原型語候補の名詞間類似度を計算し，類似度が高い略語候補と原型語候補との対応を略語と原型語との対応関係として獲得した．評価の結果，名詞間類似度の閾値を0.4に設定した場合，精度73.4%の結果を得た．また，既提案手法と比較した結果，同一の再現率においてより高い精度を達成し，既提案手法よりも有効な手法であることを確認した．本手法と既提案手法とでは，名詞間類似度の計算法が異なる．既提案手法では略語候補，原型語候補が出現する文に含まれる全ての名詞を取得し，それを意味属性に汎化する．そして，その意味属性に付与された重みを要素としたベクトルを生成する．しかし，略語候補や原型語候補と関連が低い名詞の意味属性をも要素とし，多くの非零要素を持つベクトルが生成される．それに対して，本手法は，略語候補や原型語候補に対するベクトルを生成する際に略語候補や原型語候補に関連のある名詞を抽出する処理を行い，無関係な名詞によるベクトル成分への悪影響を抑制することができた．

謝辞 本研究の一部は，文部科学省科学研究費特定領域研究(B)(2)16092213，及び，２１世紀COEプログラム「インテリジェントヒューマンセンシング」(豊橋技術科学大学)の援助により行われた．また，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

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley.

Barzilay, R. and Mckeown, K. R. (2001). "Extracting paraphrases from a parallel corpus." In *Proceedings of ACL2001*, pp. 50–57.

Hindle, D. (1990). "Noun classification from predicate argument structures." In *Proceedings of the 28th Annual Meeting of ACL*, pp. 1268–1275.

Hisamitsu, T. and Niwa, Y. (2001). "Extracting useful terms from parenthetical expressions by combining simple rules and statistical measures: A comparative evaluation of bigram statistics." In *Recent Advances in Computational Terminology, Edited by D. Bourigault, C. Jacquemin, M. -C. L'Homme, John Benjamins Publishing Company*, pp. 209–224.

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) (1997). 日本語語彙大系. 岩波書店.

乾健太郎 藤田篤 (2004). "言い換え技術に関する研究動向." 自然言語処理, **11** (5), 151–198.

石井直樹, 平石智宣, 延澤志保, 斎藤博昭, 中西正和 (2000). "日本語略語の自動復元." 情報処理学会研究報告SIG-SLP, pp. 23–30.

片岡明, 増山繁, 山本和英 (2000). "動詞型連体修飾表現の "N1 のN2" への言い換え." 自然言語処理, **7** (4), 79–98.

近藤恵子, 佐藤理史, 奥村学 (1999). "「サ変名詞＋する」から動詞相当句への言い換え." 情報処理学会論文誌, **40** (11), 4064–4074.

Lin, D. (1998). "Automatic retrieval and clustering of similar words." In *Proceedings of the COLING-ACL*, pp. 768–774.

野呂康洋, 桝井文人, 河合敦夫 (2003). "WEB 文書中の間接共起情報を利用したカタカナ語省略形の推定." 2003 年度電気関係学会東海支部連合大会講演論文集, p. 257.

Pereira, F., Tishby, N., and Lee, L. (1993). "Distributional clustering of English words." In *Proceedings of the 30th Annual Meeting of the ACL*, pp. 183–190.

Rowe, N. C. and Laitinen, K. (1995). "Semiautomatic disabbreviation of technical text." *Information Processing & Management*, **31** (6), 851–857.

酒井浩之 増山繁 (2002). "名詞とその略語の対応関係のコーパスからの自動獲得." 電子情報通信学会論文雑誌, **J85-D-II** (10), 1624–1628.

Salton, G. (1988). *Automatic Text Processing*. ADDISON WESLEY.

Terada, A. and Tokunaga, T. (2001). "Automatic disabbreviation by using context information." In *Proceedings of NLPRS2001 Post-Conference Workshop on Automatic Paraphrasing: Theories and Applications*, pp. 21–28.

真田亜希子, 舟宝貴志, 梅村恭司, 山本英子 (2003). "シソーラス自動構築システムの性能向上

のための実験.” 「情報アクセスのためのテキスト処理」シンポジウム発表論文集, pp. 120–127.

略歴

酒井 浩之: 2002年 豊橋技術科学大学大学院工学研究科修士課程 知識情報工学専攻 修了．2005年 豊橋技術科学大学大学院工学研究科博士後期課程 電子・情報工学専攻 修了．2005年 豊橋技術科学大学知識情報工学系助手．博士(工学)．自然言語処理，特に，テキスト自動要約の研究に従事．言語処理学会，人工知能学会各会員．e-mail: sakai@smlab.tutkie.tut.ac.jp

増山 繁: 1977年 京都大学工学部数理工学科卒業．1982年 同大学院博士後期課程単位取得退学．1983年 同修了(工学博士)．1982年 日本学術振興会奨励研究員．1984年 京都大学工学部数理工学科助手．1989年 豊橋技術科学大学知識情報工学系講師．1990年 同助教授．1997年 同教授．2005年 豊橋技術科学大学インテリジェントセンシングシステムリサーチセンター教授 併任．アルゴリズム工学，特に，並列アルゴリズム等，及び，自然言語処理，特に，テキスト自動要約等の研究に従事．言語処理学会，電子情報通信学会，情報処理学会等会員．e-mail: masuyama@tutkie.tut.ac.jp

(2005 年 1 月 10 日 受付)
(2005 年 3 月 29 日 再受付)
(2005 年 5 月 10 日 採録)