コーパスからの名詞と略語の対応関係の自動獲得

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

本稿では，ある名詞 (複合名詞を含む) と同一の意味をもつ略語をコーパスから自動的に獲得し，換言事例として収集する試みについて述べる．一般に，換言とは「ある表現を意味内容を保ったまま別の表現に変換する」ことである (4)．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換言事例が多く存在する．例えば，「原子力発電所」を「原発」と換言するような事例である．このような事例を本稿では略語と定義し，略語と，それに対応する名詞との対応を新聞記事から自動的に収集することを目的とする．略語と名詞の対応関係を取得できれば，あいまい性解消や，検索，要約の処理において有用である．

関連研究として，括弧表現から換言事例を自動抽出する試みがある (2)．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」を「北朝鮮」とする換言事例を抽出している．この手法により本稿で述べる手法では抽出できない換言事例を抽出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表現しか抽出できない．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」の対応は抽出できない．しかし，その略語が一般的に知られる表現であれば，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」を「原発」と換言する事例を抽出できる．また，文献 (1) では，異なった翻訳者によって翻訳された文書を比較することで，換言事例を抽出している．しかし，この手法ではパラレルコーパスとして使用した，異なった翻訳者によって翻訳された文書に出現する換言事例しか抽出できない．略語と名詞の対応関係の取得の先行研究としては，一方は略語を多く含み，もう一方は略語を含まない同一ドメインの文書群を2つ用意して，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試みがある (6)．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しか抽出できない．

これらの先行研究に対して，我々は，単一のコーパスから名詞と略語との対応を抽出する試みを以前に行なった (5)．この手法は，日本語における名詞の文字情報を用いて，ある名詞の略語である可能性のある名詞をコーパスから探索し，探索された名詞が略語であるかどうかは，名詞の共起情報から重みを算出し判定する．我々の既提案手法 (5) 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，略語を抽出可能である．しかし，我々の既提案手法は，再現率の点で改善の余地があった．本タスクの性能評価において，知識の精度は重要な要因であるが，知識の再現率も重要な要因である．なぜなら，再現率が高くないと，その手法によって獲得された知識を要約や検索に使用しても高い効果が期待できないからである．とくに，コーパス中において頻出する略語と名詞の対応関係を高い再現率で獲得することは，知識の応用性の観点から重要なことである．本稿で提案する手法は，我々による既提案手法と同等の精度とより高い再現率を目指したものであり，評価の結果，既提案手法より高い精度，再現率を達成した．

2 略語の抽出手法

本研究では，語形の一部を省略し短縮した語を略語と定義する．例えば，「原子力発電所」の「原子力」から「原」以外を省略し，「発電所」から「発」以外を省略して形成された「原発」は略語である．また，複合名詞から一部の情報が欠落しても意味が変わらない，あるいは，読み手が欠落した情報を補完できる名詞が存在する．例えば「バブル経済崩壊」から「経済」が欠落して「バブル崩壊」と表記されても一般知識から「経済」を補完することができる．この対応は，要約や検索において有用な知識であるため，本手法では，そのような名詞の対応も抽出対象とする．そのため，本稿では，そのような複合名詞から一部情報が欠落した名詞も同じく略語と定義する．

2.1 略語可能性名詞の取得

本手法は，あるコーパスに含まれている名詞を比較して名詞と略語の対応を判定し，略語を抽出する．しかし，日本語における略語は数多く存在するため，まず，ある名詞の略語である可能性のある名詞を，名詞の文字情報を用いて判別する．以降，ある名詞の略語である可能性のある名詞を**略語可能性名詞**と定義する．本稿では，略語可能性名詞とは以下のような条件を全て満たす名詞 A であると定義する．ここで，名詞 A を構成する文字を全て含み，かつ，その出現順序が等しい名詞 B の集合を $P(A)$ とする (つまり，$B \in P(A)$)．そのような名詞が存在しないとき，$P(A) = \emptyset$ である．

条件 1 名詞 $B \in P(A)$ には名詞 A が含まれていない．

条件 2 名詞 A と名詞 $B \in P(A)$ の先頭の文字が同一である．

例えば名詞 A「原発」は，名詞 $B \in P(A)$「原子力発電所」に対して全ての条件を満たす．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．しかし，「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．さらに，両名詞の先頭の文字は共に「原」であり，同一である．この条件を満たす名詞 A は略語可能性名詞であり，それに対応する名詞 $B \in P(A)$ を**原型名詞**と定義する．なお，この 2.1 節の手法は既提案手法 (5) と同一である．しかし，これらの条件だけでは，原型名詞が「三菱重工業」で，それに対する略語可能性名詞が「三重」というような明らかに間違った対応関係も抽出してしまう．そこで，略語可能性名詞と原型名詞の類似度を計算し，正しい対応関係を判定する．なお，以降は既提案手法と異なる手法である．相違点については 4.2 節で述べる．

2.2 略語と原型名詞の対の判別

略語可能性名詞と原型名詞の対に対して，略語と原型名詞の対応であるかどうかを判別するために，ベクトル空間法 (3) を元にした処理を行なう．具体的には，略語可能性名詞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可能性名詞と関連のある文書を抽出する．次に，その文書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可能性名詞と関連が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，略語可能性名詞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．同様の処理を行ない，原型名詞と関連がある名詞を抽出し，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とり，余弦がある閾値以上の略語可能性名詞と原型名詞の対応を略語と原型名詞の対応と判定する．以下にアルゴリズムを示す．

略語の判定アルゴリズム

Step 1 略語可能性名詞 A を含む文書集合 $D(A)$ を抽出する．

Step 2 文書集合 $D(A)$ の各文書 $d \in D(A)$ に対して，式 (1) によって計算される重みを付与する．

$$W(d) = \frac{tf(A,d)}{1+\log(\sum_{t\in T(d)} tf(t,d))} \cdot \frac{1+nl(d)-nlf(A,d)}{nl(d)} \quad (1)$$

但し，

$tf(A,d)$: 文書 $d \in D(A)$ において，略語可能性名詞 A が出現する頻度，

$T(d)$: 文書 $d \in D(A)$ に含まれる名詞の集合，

$nl(d)$: 文書 $d \in D(A)$ の行数，

$nlf(A,d)$: 文書 $d \in D(A)$ において，略語可能性名詞 A が最初に出現する行，

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 $D_n(A)$ とする．

Step 4 文書集合 $D_n(A)$ に出現する名詞の集合 $T(D_n(A))$ に含まれる各名詞 $t \in T(D_n(A))$ に対して，式 (2) によって計算される重みを付与する．

$$W(t) = \frac{TF(t,D_n(A))}{1+\log(\sum_{s\in T(D_n(A))} TF(s,D_n(A)))} \cdot \log\frac{|N|}{df(t,N)} \quad (2)$$

$$TF(t,D_n(A)) = \sum_{d\in D_n(A)} tf(t,d)$$

但し，

N: 対象としたコーパスの文書集合，

$df(t,N)$: 対象コーパスにおいて，名詞 $t \in T(D_n(A))$ が出現する文書数，

Step 5 Step 4 によって付与された重みが大きい上位 m 名詞を抽出する．

Step 6 Step 5 によって抽出された名詞の重みを要素としたベクトル $V(A)$ を生成する．

Step 7 原型名詞 B についても Step 1 から Step 6 の処理を行ない，ベクトル $V(B)$ を生成する．

Step 8 2 つのベクトル $V(A)$ と $V(B)$ の余弦をとり，余弦がある閾値以上の略語可能性名詞と原型名詞の対応を略語と原型名詞の対応と判定する．

□

なお，Step 4 における名詞 t が複合名詞である場合，その複合名詞自身と，その中に含まれる普通名詞が文書集合 $D_n(A)$ に出現しているとする．例えば，複合名詞「株式市場」なら「株式市場」「株式」「市場」の 3 つの名詞が出現していることになる．また，名詞 t が「こと」である場合と数詞である場合は，それらを排除する．

2.3 省略名詞における各種制限

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，原型名詞に含まれる普通名詞の中で，略語を構成する文字を含まない名詞と定義する．例えば，略語「バブル崩壊」と，その原型名詞「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．しかし，略語「原発」と，その原型名詞「原子力発電所」では，省略名詞は存在しない．なぜなら,「原子力発電所」には,「原子力」「発電所」の 2 つの普通名詞が含まれるが,「原発」を構成する文字は 2 つの普通名詞に共に含まれているからである．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 もし，原型名詞の中に省略名詞が複数存在していれば，その略語可能性名詞と原型名詞の対応を略語と原型名詞の対応と判定しない．

制限 2 原型名詞の最後の普通名詞 (例えば，原型名詞「バブル経済崩壊」なら「崩壊」) が省略名詞であるならば，その略語可能性名詞と原型名詞の対応を略語と原型名詞の対応と判定しない．

制限 1 について説明する．省略名詞が存在している原型名詞が，対応する略語可能性名詞と意味が同等であると判断できるのならば，その省略名詞が補完できなければならない．しかし，省略名詞が複数存在しているならば，それらを全て補完できる可能性は低い．そのため，制限 1 を加える．

制限 2 について説明する．原型名詞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．

3 手法の実装

本手法を実装して略語の抽出を行なった．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 84905 記事を採用した．形態素解析器として JUMAN [1] version 3.5 を採用した．また，上位 n 文書の n を 20，上位 m 名詞の m を 200 とした．本手法を実装したシステムによって取得した略語と原型名詞の対応には，「原子力発電所」に対する略語「原発」や，「生命保険」に対する略語「生保」があった．

4 評価実験

4.1 評価方法

本手法を精度，再現率で評価する．まず，正解データを作成する必要があるが，ある名詞の略語，もしくは，略語に対する原型名詞を人手によって作成するのは困難な作業である．なぜなら，略語と意識せずに使用している名詞や，原型名詞に対する略語が存在しているかどうか分からないことが多いからである．そこで，本手法の正解データは，2.1 節で述べた方法で取得した略語可能性名詞と原型名詞の対応を人手で判定することで作成する．具体的な評価方法を以下に示す．

評価実験方法

Step 1 略語可能性名詞 A と原型名詞 B の対応に対して，以下の式でコーパスにおける A と B の出現確率の調和平均 $F(A,B)$ を求める．なお，N は対象としたコーパスの文書集合，$df(t,N)$ は，文書集合 N において名詞 t を含む文書数である．

$$F(A,B) = 2P(A)P(B)/(P(A)+P(B))$$
$$P(t) = df(t,N)/|N|$$

Step 2 $F(A,B)$ が高い値をもつ上位 1000 の略語可能性名詞 A と原型名詞 B との対応を抽出する．

[1] http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

表 1: 本手法の精度，再現率

閾値	精度 (%)	再現率 (%)	抽出数
0.2	64.3	68.9	210
0.3	72.0	59.2	161
0.4	73.9	44.9	119
0.5	76.4	34.7	89

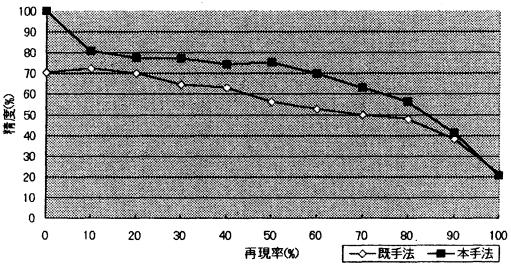

図 1: 既提案手法との比較評価

Step 3 抽出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，精度，再現率を計算する．

□

4.2 比較手法

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 (5) と同等の精度とより高い再現率を目指したものである．本手法と既提案手法 (5) とは，2.1 節で述べた略語可能性名詞の抽出手法は同一である．また，ベクトル空間法を応用した部分も同一であるが，ベクトルを生成する要素が異なる．本手法は，略語可能性名詞 (もしくは原型名詞) が出現する文書に含まれる名詞の重みを要素としてベクトルを生成する．その際，略語可能性名詞 (もしくは原型名詞) 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．

それに対して，既提案手法では略語可能性名詞 (もしくは原型名詞) が出現する文に含まれる全ての名詞を意味素性に汎化する．そして，その意味素性に付与された重みを要素としたベクトルを生成する．既提案手法では，略語可能性名詞 (もしくは原型名詞) と関連のない名詞を排除する処理を行なっていない．また，2.3 節の制限は本手法のみに課してあり，既提案手法では課していない．

4.3 実験結果

本稿で提案した手法と比較手法とを，精度，再現率で評価する．本手法と比較手法を実装し，それぞれの精度，再現率を計算する．図 1 に，再現率を X 軸とした場合の精度の変化を示した実験結果を示す．また，表 1 に閾値を 0.2 から 0.5 まで変化させた場合の本手法の精度，再現率，抽出数を示す．

5 考察

図 1 から，本手法は再現率を保ちつつも精度が低下しておらず，既提案手法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60%まで精度70%を維持している．しかし，既提案手法は，再現率が20%までなら精度70%を維持しているが，再現率が高くなると，精度も大きく低下してしまい，高い精度における再現率が低い．その理由として以下が考えられる．

既提案手法は，略語可能性名詞 (もしくは原型名詞) が出現する文に含まれている全ての名詞の意味素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ベクトル間の共有要素数が少ないため，それらを全て採用し，さらに名詞を意味素性に汎化する必要があった．そのため，略語可能性名詞 (もしくは原型名詞) と関連が低い名詞の意味素性をも要素とし，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素性は，略語可能性名詞と原型名詞とで共有要素となる確率は低いが，もし，略語可能性名詞と原型名詞の対応が正解であった場合に余弦の値が小さくなる原因になる．そのため，既提案手法は，本来，正解である略語可能性名詞と原型名詞の対応を正解と判定できず，再現率が低下したものと考える．

それに対して，本手法は略語可能性名詞 (もしくは原型名詞) 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可能性名詞 (もしくは原型名詞) 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，略語可能性名詞 (もしくは原型名詞) 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，本手法は略語可能性名詞 (もしくは原型名詞) に対するベクトルを生成する際に，略語可能性名詞 (もしくは原型名詞) 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影響を抑制することができた．そのため，高い精度における再現率の低下を抑えることができたと考える．また，既提案手法は意味素性辞書の精度に左右されてしまうが，本手法は，意味素性辞書を必要としないのも利点である．

本手法は，略語可能性名詞や原型名詞に「ドル」や「円」といった，通貨の名称がついている場合に多く誤判定を行なった．また,「前期比」といった経済指標を表す用語で，多く誤判定を行なった．例えば原型名詞「ドイツマルク」に対する略語可能性名詞「ドル」を略語と判定した．2つの名詞は意味は異なるが，通貨の名称という点で関連性がある．本手法は，略語可能性名詞 (もしくは原型名詞) と関連が高い名詞を要素とするベクトルの余弦で判定を行なうため，2つの名詞の関連性が高いと要素となる名詞にも共通の名詞が出現し，余弦が大きくなってしまう．そこで，略語可能性名詞や原型名詞に「ドル」,「円」,「比」を含む対応を略語と原型名詞の対応と判定しないという規則を課した．その場合の精度，再現率を表2に示す．表1の結果と比べると，再現率はほとんど低下していないにもかかわらず，精度が向上していることが分かる．このような規則はアドホックなものであり，日本経済新聞というコーパスの性質に特化しているが，精度の向上に有効な規則である．このような規則を追加することで，精度の向上を行なうアプローチがあると考える．また，関連性が高い略語可能性名詞と原型名詞の対応を，なんらかの手法で判定対象から除外する必要があるが，その手法によって再現率の低下を招く可能性もある．この問題は今後の課題とする．

表 2: 本手法の精度，再現率 (規則あり)

閾値	精度 (%)	再現率 (%)	抽出数
0.2	68.7	66.3	189
0.3	77.6	56.6	143
0.4	78.8	43.9	109
0.5	80.5	33.7	82

6 まとめ

本稿では，ある名詞と同一の意味をもつ略語をコーパスから自動的に獲得し，換言事例として収集する試みを行なった．本手法では，名詞とその略語候補に対して，それぞれ，文書における名詞の共起情報から重みを算出し，その名詞の重みを要素としたベクトルを作成する．そして，2つのベクトルの余弦で名詞と略語の対応を判定する．本手法の評価を行ない，いくつかの規則を課した場合，再現率56.6%において精度が77.6%を得ることができ，既提案手法よりも良好な結果を得た．

謝辞

本研究は文部科学省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」，及び，日本学術振興会科学研究費基盤研究 (C)(2)13680444 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．

References

[1] R. Barzilay, K. R. Mckeown, "Extracting Paraphrases from a Parallel Corpus," Proc. of ACL2001, pp. 50-57, 2001.

[2] T. Hisamitsu, Y. Niwa, "Extracting useful terms from parenthetical expressions by combining simple rules and statistical measures," Didier Bourigault, Christian Jacquemin and Marie-Claude L'Homme(eds.), Recent Advances in Computational Terminology, John Benjamins Publishing Company, pp. 209-224, 2001.

[3] G. Salton, "Automatic Text Processing," Addison-Wesley, 1988.

[4] 佐藤理史，"論文表題を言い換える，" 情報処理学会論文誌，vol.40，no.7，pp. 2937–2945，1999.

[5] 酒井浩之，増山繁，"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，vol.J85-D-2，no.10，pp. 1624–1628，2002.

[6] A. Terada, T. Tokunaga, "Automatic disabbreviation by using context information," Proc. NLPRS2001 Post-Conference Workshop on Automatic Paraphrasing:Theories and Applications, pp. 21–28, 2001.