

# 原語における形態素間の相対的な優先関係を利用した 略語自動推定

及川 孝徳<sup>†1</sup> 山名 早人<sup>†2</sup> 大西 貴士<sup>†3</sup> 土田 正明<sup>†3</sup> 石川 開<sup>†3</sup>

<sup>†1</sup> 早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1

<sup>†2</sup> 早稲田大学理工学術院 〒169-8555 東京都新宿区大久保 3-4-1

<sup>†3</sup> NEC 情報・メディアプロセッシング研究所

E-mail: <sup>†1</sup>, <sup>†2</sup> {t\_oikawa,yamana}@yama.info.waseda.ac.jp

<sup>†3</sup> {t-onishi@bq,m-tsuchida@cq,k-ishikawa@dq}.jp.nec.com

あらまし インターネットの普及に伴い、ウェブ上に存在する情報は増え続け、特定の対象に関する情報を網羅的に収集する技術が注目されている。特定の対象に関する情報を収集する場合、対象の正式名称だけでなく同義語での検索も行わなければ、網羅的に文書を収集することはできない。そのため、対象の同義語を自動推定する技術が求められる。本稿では、同義語の一つである略語に関する手法を提案する。与えられた原語から略語を自動推定するこれまでの研究は、略語候補として原語からすべての部分文字列を生成し、候補に順位付けを行い、上位から出力する手法を用いている。候補の順位付け方法には、原語から略語になるプロセスを利用する研究や音や拍数の変化に着目した研究がある。しかし、既存研究は原語の表層的な情報のみを利用しており、原語を構成する形態素同士の優先関係に着目していない。このため、形態素の意味内容に基づいた情報が活かされていない。これに対して本稿では、従来の研究が扱ってこなかった形態素間の関係に着目し精度及び網羅性の向上を試みる。具体的には、原語における形態素間の相対的な優先関係をペアワイズにより判定する手法を提案する。提案手法は、2段階から構成される。1段階目では、形態素同士を比較し相対的な優先関係をみることで、原語から略語に用いる形態素を選択する。2段階目では、選択された形態素を統計的に求めた言語的省略ルールによって変形する。

キーワード 略語推定, クエリ拡張

## Automatic Estimation of Abbreviated Words Using Relative Priority Relation between Morpheme

Takanori OIKAWA<sup>†1</sup> Hayato YAMANA<sup>†2</sup>

Takashi ONISHI<sup>†3</sup> Masaaki TSUCHIDA<sup>†3</sup> Kai ISHIKAWA<sup>†3</sup>

<sup>†1</sup> Graduate School of Fundamental Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555, Japan

<sup>†2</sup> Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555, Japan

<sup>†3</sup> NEC Information Media Processing Reserch

E-mail: <sup>†1</sup>, <sup>†2</sup> { t\_oikawa,tai,matuzaki,kuroki,yamana }@yama.info.waseda.ac.jp

<sup>†3</sup>, t-onishi@bq.jp.nec.com,m-tsuchida@cq.jp.nec.com,k-ishikawa@dq.jp.nec.com

**Abstract** No English .

**Keyword** Estimation for Abbreviated Words, Query Extraction

### 1. はじめに

インターネットの普及に伴い、ウェブ上に存在する

情報は増え続けている。ウェブ上の文書情報が増えるにつれ、大量の文書から特定の人物や商品について書かれた情報を獲得する技術へのニーズが、企業・一般

ユーザー問わずに高まっている．特定の対象について書かれた文書を収集する場合を考えると，対象の正式名称だけでなく，対象の同義語での検索も行わなければ，網羅的に文書を収集することはできない．そのため，対象の同義語を自動的に推定する技術が求められる．

本論文では，対象の同義語を自動的に推定する技術の一つである略語推定の精度及び再現率を向上させる手法について提案する．略語とは，正式名称から任意の文字を削除することで作成される同義語の一種であり，新聞やメモ書きといった，字数を制限したい文書や，共通認識を持った特定コミュニティ内のユーザー間会話等で用いられる．本稿では、『略語』に対し，省略される前の正式名称を『原語』と呼ぶ．英語の略語のほぼすべてが，原語を構成する単語の頭文字をとることで生成されるのに対し，日本語の略語生成規則は複雑であるため，日本語の原語から略語を推定するには，工夫が必要になる．

略語を推定する研究は，括弧表現等の原語－略語関係を示す特定表現を利用した研究[1][2]，コーパス中に存在する名詞の中から，出現状況の類似を見て原語－略語の関係対を抽出する研究[3][4]，与えられた原語から略語を自動推定する研究[5][6][7][8]の 3 つに分類できる．最初の 2 つの研究は，与えられたコーパスから略語辞書を作成することを目的としている．対して，与えられた原語から略語を自動推定する研究は，コーパスに縛られることなく，指定した原語の略語を推定することができる．本稿では，これらの分類の内，3 番目の「与えられた原語から略語を自動推定する」ことを目指す．

与えられた原語から略語を自動推定する分野の従来研究は，略語候補として原語からすべての部分文字列を生成し，候補に順位付けを行い，上位から出力する手法を用いている．順位付けの方法は，原語と略語候補をそれぞれ含む文書集合の類似度を使う手法[5]，原語から略語が生成されるプロセスモデルを使う手法[6][7]，モーラ・シラブルといった音の特徴を使う手法[8]がある．しかし上記の手法は，文字の表層的な情報のみを使っており，原語を構成する形態素同士の優先関係は用いていない．また，略語候補をすべて生成し，候補に順位付けする手法は，原語が一文字長くなるごとに 2 の階乗に計算量が増えるため，順位付けにかかるコストによっては計算量が膨大になってしまう．これに対して，形態素間の優先関係，すなわちある形態素 A と別の形態素 B が原語に含まれた時、どちらが省略されやすいかといった情報を用いることができれば，精度，再現率の向上ができると共に候補となる略語数を減らすことができる．

そこで本稿では，原語における形態素間の相対的な優先関係を利用することで，略語自動推定の精度・網羅率向上，さらには略語候補数の削減を目標とした．相対的な優先関係とは 2 つの形態素を比較したときに，どちらが情報として重要かという尺度である．提案手法は，略語の自動推定を、『原語情報の取捨選択』と『略語らしい形への変形』の 2 段階に分けて行う．1 段階目では，原語の形態素同士を比較し，相対的な優先関係によって略語に使用する形態素の選択を行う．提案手法においては，文字の表記情報，品詞情報，漢字仮名情報を，優先関係を比較する要素として使用した．2 段階目では，1 段階目で得られた形態素を言語的省略ルールに基づき，略語らしくなるよう変形する．ここで言われる「略語らしさ」は略語の文字長を意味しており，各形態素を言語的省略ルールによって変形しつつも，全体におけるバランスを調整していることを示す．言語的省略ルールを用いることで，存在し得ない候補の生成を防ぎ，候補数が膨大になるのを避けることができる．最後に，1 段階目・2 段階目それぞれで付加されたスコアによって順位付けを行い，上位から略語候補を出力する．以上の手法によって，略語を自動推定する上で，精度と再現率の向上が可能となる．

本稿の構成は以下のとおりである．第 2 節では関連研究について述べる．第 3 節では提案手法について説明する．第 4 節では提案手法の評価実験について述べ，第 5 節でまとめを述べる．

## 2. 関連研究

### 2.1. 特定表現を利用した略語抽出

久光ら[1]は新聞記事における括弧表現を利用し，略語の抽出を行った．新聞記事における高頻度の括弧表現を，括弧内外の関係により，「A(B)」を以下の 4 つに分類し，略語の抽出を行なっている．

- A または B が英数字か片仮名のみ  
→ 言い換え
- B が平仮名のみ  
→ 読み
- A, B ともに年号  
→ 言い換え
- B の長さが 3 以上かつ，B に含まれる語の半分以上が順序を保って A に含まれる  
→ 言い換え（略称）

括弧表現を効果的に獲得するため，A・B の共起強度を統計的指標として用いており，括弧表現全体に含まれる原語－略語対の約 3 分の 2 が抽出可能としている．

岡崎ら[2]は久光ら同様括弧表現に着目し，括弧表現

において言い換えの関係が成立するか調べる指標として言い換え発生率を定義している．言い換え発生率は以下の2つの条件を満たす文書の割合で表される．

- 「A(B)」のパターンが出てくる前の文において、表現 B が出現しない
- 「A(B)」のパターンが出てきた後の文において、表現 A よりも表現 B の出現頻度が高い

2.2. コーパス中での出現状況から略語を抽出する手法

酒井ら[3]は、特定コーパス中の名詞同士を比較し、原語と略語のペアであるかの判定を行なっている．

まず、略語候補としてある原語を構成する文字を順序どおり含む名詞を挙げる．次に、原語を含む文書集合と略語候補を含む文書集合中の名詞をすべて抽出し、日本語語彙体系によって付与した意味素性のコサイン類似度によって判定を行なっている

後に酒井らは改良を行っており[4]、略語候補を含む文書集合、原語を含む文書集合に含まれる名詞を取得し、名詞の重みを要素としてベクトルを生成している．ベクトル生成の際、略語候補・正式名称の双方に関連度の低い名詞の除去処理を行なっている．

2.3. 原語から略語を自動推定する研究

原語から略語を自動推定する研究では、原語文字列から任意の文字を抜くことで生成できるすべての略語候補を生成し、候補に独自の指標で順位付けを行い、出力する手法を用いている．

関口ら[5]は原語を含む文書集合と、略語候補を含む文書集合の内容の類似度によって、略語候補の順位付けを行った．

村山ら[6][7]は、形態素の削除とその位置、文字の削除のされ方といった、原語から略語が生成されるプロセスをモデル化し、略語候補の順位付けを行なっている．モデルは、形態素選択と各形態素変形それぞれに作成している．

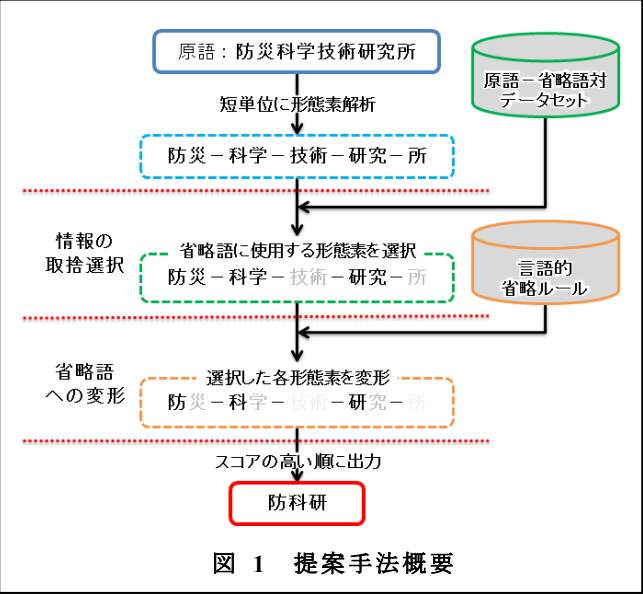
和田ら[8]は、人間の感覚を考慮し、モーラ・シラブル（拍と音節）をCRF（条件付き確率場）の素性に利用し、略語候補の順位付けを行った．

上記の3つの手法は、原語から想定される略語候補をすべて列挙し、それぞれ独自の指標によって順位付けする手法をとっている．ただし、村山ら、和田らの手法は、最終的にウェブ情報を用いて順位を修正した後に推定結果を出力している．

3. 提案手法

提案手法では、原語における形態素間の相対的な優先関係を利用することで、略語自動推定の精度・網羅率向上、さらには略語候補数の削減を目標とした．こ

れは、従来の手法が、文字の表層的な情報のみを用いているのに対し、原語を構成する形態素同士の優先関係を用いることによって、精度及び再現率の向上ができると共に候補となる略語数を減らすことができると考えたからである．提案手法の流れを図1に示す．



提案手法では、略語自動推定の処理を、『原語情報の取捨選択』、『略語らしい形への変形』の2段階に分けている．2段階に分けることによって、省略語生成候補を効率的に減らすことができる．

まず、前処理として与えられた原語を国立国語研究所[9]が定めた短単位に形態素分析する．1段階目の、『原語情報の取捨選択』では、原語－略語対データを用いて、与えられた原語中の形態素から、略語候補生成に使う形態素を選択する．2段階目の、『略語らしい形への変形』では、1段階目で得られた形態素を、言語的ルールによって、略語らしくなるよう変形する．

3.1. 原語情報の取捨選択

原語－略語対データを元に、略語候補に使う形態素を選択する．まず、各形態素に対し、先頭であるかどうか、漢字仮名、品詞の属性を付加する（表1）．

| 防災               | 科学                 | 技術               | 研究                 | 所                |
|------------------|--------------------|------------------|--------------------|------------------|
| 先頭               | －                  | －                | －                  | －                |
| 漢字               | 漢字                 | 漢字               | 漢字                 | 漢字               |
| 名詞<br>普通名詞<br>一般 | 名詞<br>普通名詞<br>サ変可能 | 名詞<br>普通名詞<br>一般 | 名詞<br>普通名詞<br>サ変可能 | 接尾辞<br>名詞的<br>一般 |

表1 各形態素への属性付加

次に、形態素のペア間で優先度（略語化される際の残りやすさ）の比較を行う．具体的には、対象形態素*i*と比較している形態素*j*において、2つが含まれる原語の略語において、形態素*i*だけが選択される確率をスコアとして求める（式(1)）．

$$\text{Score}(i) = \left\{ \frac{NP'(i)}{NP(i)} \right\} \times \sum_{j \in O(i)} \left\{ \frac{NP'(i) - 0.5 * NP'(i \cap j)}{NP(i \cap j)} \right\} \quad (1)$$

式(1)において、P は学習データの原語集合、P'は略語集合を示し、NP(i)は P 中において形態素 i を含む原語の数、NP'(i)は P'中において形態素 i を含む略語の数を示す。また、O(i)は対象原語に含まれる形態素 i 以外の形態素集合である。

同様に、他の属性に関してもスコアを計算し、その和を各形態素の重要度と定義する。

### 3.2. 略語らしい形への変形

3.1 節で得られた形態素を、言語的ルールによって略語らしい形へと変形する。提案手法では、変形パターンを、形態素が漢字であるときと仮名であるときで処理を変えている。ルールは、[13]と原語－略語対データセットを用いて統計的に作成している。例えば形態素がカタカナで構成されている場合は図 2 のルールを基準としている。

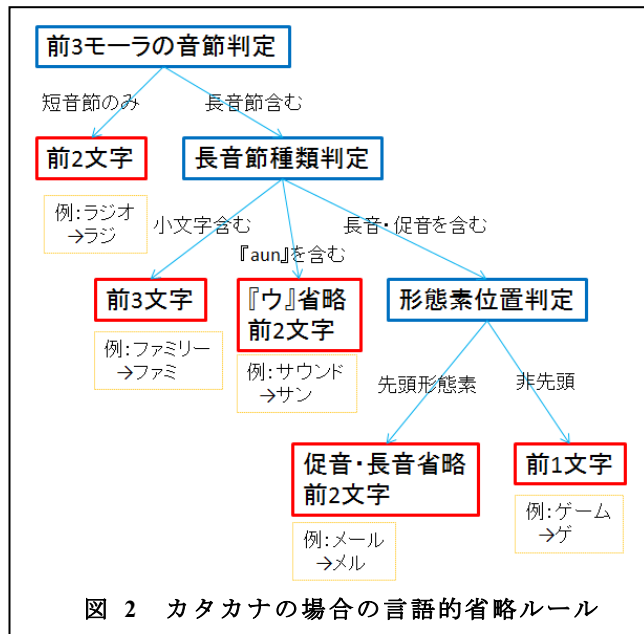


図 2 カタカナの場合の言語的省略ルール

基準となる変形に重みをおきつつ、他の変形パターンも生成する。

各形態素変形スコアを、選択した形態素数で除算した値を略語候補のスコアとする。

### 3.3. 略語候補にスコアをつけ出力

3.1 節、3.2 節で得られたスコアそれぞれに重みをつけた上で加算し、略語候補の最終的なスコアとする。

## 4. 評価実験

### 4.1. 使用データ・実験環境

#### ➤ 原語－略語対データ

原語－略語対データは村山ら[6][7]の手法で用いて

いた方法を参考に、Wikipedia[10]と Yahoo ウェブ検索[15]を利用し作成を行った。手順を以下に示す。

- ① Wikipedia の項目データを収集する
- ② ①で得た項目をクエリとし、以下のテンプレートに当てはめたもので、ウェブ検索を行う。

表 2 テンプレートと略語の出現位置

| テンプレート    | 略語の出現位置 |
|-----------|---------|
| 「正式名称」の略  | 前後      |
| 「正式名称」略して | 後       |
| 「正式名称」略称  | 前後      |

- ③ 取得したスニペット中「略語出現位置」から、正式名称の文字長分（記号を除く）を抽出。
- ④ 抽出した部分スニペット中で正式名称の文字列の一部を、正式名称と同じ順に含む部分を、略語候補として抽出。
- ⑤ 取得した部分スニペットすべてにおける出現回数で略語候補の足切りを行う。
- ⑥ 得られた略語候補が正しいかを人手で判断する。今回は学生 3 人のうち 2 人が正しいとしたものを正解と判断した。

この方法で得た正解セットに、Wikipedia[10]の「大学の略称」ページと、証券コード辞書[14]を利用して、原語－略語対データを作成した。

#### ➤ 言語的省略ルール

言語的省略ルールは、『新語はこうして作られる』[13]を参考に、原語－略語対データを統計的・経験的に解析し、作成した。

#### ➤ 実験環境

実験環境として、形態素解析には茶筌[11]を利用し、形態素解析のレベルは国立国語研究所[9]の定める短単位を選択した。

#### ➤ 評価セット

評価に使用する正解セットは、略して使用される機会が多いと考えられる『チェーンストア名』『漫画・アニメ』『TV バラエティ番組名』『上場企業名』の 4 ジャンルから 40 語ずつ作成した。項目名は Wikipedia[10]、証券コード辞書[14]を使用している。それぞれランダムに 25 語ずつの原語を、形態素長の偏り方が母集団と同じになるよう選択した。なお、略語を推定することが目的であるため、1 形態素からなる原語は除外している。各ジャンルの母集団原語数、形態素長分布を表 3 に示す。

表 3 評価セット形態素長分布

| 形態素長      | 2  | 3  | 4  | 5  | 6~ | 母集団の数 |
|-----------|----|----|----|----|----|-------|
| チェーンストア   | 14 | 10 | 10 | 10 | 10 | 302語  |
| 漫画・アニメ    | 5  | 7  | 4  | 4  | 20 | 5258語 |
| TVバラエティ番組 | 4  | 6  | 6  | 5  | 19 | 219語  |
| 上場企業名     | 8  | 10 | 9  | 8  | 5  | 1443語 |

抽出された原語を被験者 10 人に見せ、考えられる略語を最ももらしい順に 2 個以上記入してもらった。そして、上から 2 つまでを取得した。1 位に挙げた略語を 3 点、2 位に挙げた略語を 1 点とし、合計 3 点以上を得た略語を、正解とした。

なお、評価にあたって学習データから評価セットに含まれる原語一略語対は除去している。

#### 4.2. 評価方法

まず、形態素間の相対的な優先度の利用が、略語推定に有効か示し、次に、提案手法全体を既存手法と比較する。

#### 4.3. 評価結果

提案手法を部分的・全体的に他の手法との比較を行った。以下に結果を示す。

##### ➤ 原語情報取捨選択

提案手法における原語情報取捨選択部分について、他の手法との比較を行った。比較手法は、形態素単体の選択確率のみを利用した手法、CRF（条件付き確率場）、村山らの手法の形態素選択モデル部分の 3 つである。形態素単体の選択確率を利用した手法における、各属性値の重みづけは提案手法と同じ重みで行った。CRF ライブラリは CRF++ を使用し、素性には提案手法と同様に、表記・漢字仮名・品詞を与えた。どの手法も、学習データは同様のものを用いている。

結果を表 4 に示す。

表 4 原語情報取捨選択手法比較（全 172 語）

| 手法         | 精度(正解数)   | 精度(1位の正解率) |
|------------|-----------|------------|
| 提案手法       | 46.5%(80) | 23.3%      |
| 形態素単体の選択確率 | 35.5%(61) | 17.4%      |
| CRF        | 24.4%(42) | 13.4%      |
| 村山らの手法     | 41.8%(72) | 21.5%      |

表 4 の結果により、形態素の意味内容に基づいた情報が、略語推定に有用であることがわかる。また、形態素単体の選択確率を利用した手法より、提案手法のほうが高い精度であることから、相対的な形態素間関係を見る手法が有効であることがわかる。

##### ➤ 全体的比較

提案手法を既存研究と比較し、その優位性を示す。比較項目として、精度・再現率・学習データ量による精度変化・形態素長による精度の変化を考えている。

#### 5. まとめ

本稿では、原語における形態素間の関係を利用した、略語の自動推定の提案手法について説明し、実験を行った。

#### 参考文献

- [1] 久光徹，丹羽芳樹：“統計量とルールを組み合わせて有用な括弧表現を抽出する方法”，情処研報，NL，Vol.97，No.109，pp.113-118，1997
- [2] 岡崎直観，石塚満：“言い換え可能な括弧表現の抽出法”，言語処理学会第 13 回年次大会，pp. 911-914，2007
- [3] 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”，通学論 D-II，Vol.J85-D-II，No.10，pp.1624-1628，2002
- [4] 酒井浩之，増山繁：“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”，自然言語処理，Vol.12，No. 5，pp.207-231，2005
- [5] 関口裕一郎，佐藤吉秀，川島晴美，奥田秀範，“ブログ文書集合を用いた略語抽出手法の検討”，通信学技報，DE，No.107，pp.207-210，2007
- [6] 村山紀文，奥村学，“Noisy-channel model を用いた略語推定”，言語処理学会第 12 回年次大会，pp.837-840，2006
- [7] 村山紀文，奥村学：“Web 情報を利用した確率モデルによる略語推定”，情報研報，Vol.2008，No. 4，pp.93-100，2008
- [8] 和田健太，近山隆，横山大作，三輪誠：“素性にモーラとシラブルを用いた略語の自動推定”，情処研報，NL，Vol.2009，No.190，2009
- [9] 国立国語研究所，  
<http://www.ninjal.ac.jp/>
- [10] フリー百科事典 ウィキペディア  
<http://ja.wikipedia.org/wiki/>
- [11] 茶筌  
<http://chasen-legacy.sourceforge.jp/>
- [12] CRF++  
<http://crfpp.sourceforge.net/>
- [13] 窪園晴夫：“新語はこうして作られる”，岩波書店，2002
- [14] 証券コード辞書 201112 版  
<http://trendview.blog.ocn.ne.jp/>
- [15] Yahoo!デベロッパーネットワーク，  
<http://developer.yahoo.co.jp/>