# 漢字を中心とした複合語の略語の自動生成
## －音訓を考慮したルールを用いて－

岡田　真　　髙橋　幹浩
大阪府立大学大学院理学系研究科
okada@mi.s.osakafu-u.ac.jp

## １　序論

一般に，意味内容を保持したまま名詞を短縮して，同義語として扱うことがある．これらは一般的に略語と呼ばれている．以下，本稿では，元となった語を「原語」とし，「全ての文字が原語中に同じ順序で出現する短縮語」を「略語」と呼ぶ．略語は口語や文語のどちらでもよく使用され，原語と略語の関係を把握することは，文書検索や文書要約において非常に有用であると考えられる．

これまでの略語獲得に関する研究では，多くの場合テンプレートを用いて，原語と略語の対を獲得するものであった[1]．しかし，この手法では新たに作られた複合語などには対応しにくい．そこで，ルールを用いて略語を生成することで，前述のような複合語にも柔軟に対応できるようにする手法が考えられる．このような研究としては村山らの研究[2]が挙げられる．[1][2]どちらも，略語の読みや漢字の音訓の情報が用いられてない．

そこで本稿で提案するシステムでは，まず既存の原語と略語の関係から略語の生成ルールを取得する．そして，それらのルールを用いて，列挙された略語候補群から最適なものを出力する．その際に，生成ルールに音訓のルールを加えて使う．その有効性を実験により確かめる．なお，今回は漢字で構成されている複合語に範囲を限定する．

## ２　略語の生成ルール

ここでは，漢字で構成される複合語の略語の性質と生成ルールについて説明する．

原語から生成される略語候補は複数考えられる．しかし，同音異義語や略語を使用するグループの違いなど，さまざまな要因があるために，最適な略語を選択する方法は明確に定めることは難しい．

本稿では，以下の３つの生成ルールを用いる．

１．形態素の省略，
２．文字数の減少，
３．音訓の組合せと読み．

以下にそれぞれについて説明する．調査用の訓練データとして，文献[3]から漢字の名詞や複合語とその略語 678 組を取得した．なお，法律名のような文から作られる略語の組に関して今回は除外した．

| | 原語 | ⇒ | 形態素解析 | ⇒ | 略語 |
|---|---|---|---|---|---|
| (1) | 短期大学 | ⇒ | 短期，大学 | ⇒ | 短大 |
| (2) | 関西国際空港 | ⇒ | 関西，国際，空港 | ⇒ | 関空 |
| (3) | 伊豆急行 | ⇒ | 伊豆，急行 | ⇒ | 伊豆急 |
| (4) | 経営財団 | ⇒ | 経営，財団 | ⇒ | 営団 |

図１．形態素の省略法則の例

ルール１「形態素の省略」は，原語を形態素ごとに分けて考えたときの形態素自体の変化の法則を利用したものである．その法則は図 1 に示すように(1)先頭文字，(2)形態素非使用，(3)形態素非短縮，(4)後方文字の４種類に分けることができる．

一般的に(1)のように各形態素の先頭の文字を組み合わせて生成されることが多い．しかし，(2)のように使われない形態素がある場合，(3)のように形態素が短縮されない場合，(4)のように形態素の後方部分が使用される場合がある．形態素を多く持つ原語では，これらの法則が複数同時に適用されることも多い．ひとつの形態素が複数の省略法則をもつこともある．このように形態素の変化だけを見ても，一意的に規則を定めるのは難しい．

ただし，３文字以上の形態素においては，「中学校」が「中学」となるように，先頭もしくは後方から２字抽出されることはあるが，先頭と後方というように，中間を抜かした形で省略されるケースは確認されなかった．

ルール２「文字数の省略」は，原語の文字数と略語の文字数の関係を考慮したものである．

経験的に，略語は原語の文字数にかかわらず２文字または３文字になることが多い．短い原語の例では「水素爆弾」が「水爆」となる．長い原語の場合でも，「文部省美術展覧会」の８文字の原語から，「文展」の２文字の略語となる．実際に訓練データを用いて調査した結果でも，２文字の略語が 41.15%，は３文字の略語が 37.76%となり，あわせると２文字または３文字の略語は 78.91%となる．

ルール３「音訓の組合せと読み」では，略語のモーラ数と音訓の組合せに注目した．モーラとは拍を表す単位で，日本語の場合仮名１文字が１モーラとなる．ただし、「しゃ」「ふぁ」のような拗音は，２文字で１モーラとなる．

日本語はモーラ言語とも言われており，モーラは読みについて考える際に必要不可欠の単位である[4]．また，ひとつ大きな単位に「フット」がある．１フットは２モーラである．語句の短縮や複合語の生成など，新しい語を形成する過程において，その語は最小１フットの長さを持つとされている．ただし「手」のように，元来の読みが１モーラの語は除く．特に日本語では２モーラの長さが使われることが多く，訓練データの略語は，全ての漢字が２モーラ以下となっていた．

一部の漢字では，「名古屋駅（ナゴヤエキ）」が「名駅（メイエキ）」と略されるときの「名」のように，訓が１モーラ，音が２モーラの場合に，略語が生成される段階で訓から音に読みが変化することがある．同様の理由で音から訓に読みが変化する例もある．また，音訓どちらも２モーラのときに，重箱読みや湯桶読みを回避するために，音を優先して読みが変化することが多い．

これらのことから，読みの評価をするためには，略語のモーラ数を得るだけではなく，略語を構成する漢字の音訓情報を取得する必要性がある．

# 3　略語生成システム

## 3.1　システムの概略

本稿での提案システムでは，原語から略語の定義に沿う略語候補全てを列挙し，それらの候補の略語らしさについて，生成ルールを用いて評価値を算出し，その値の和が大きな候補を最適な略語として出力する．システムの流れは以下の手順となる．

1. 略語の定義に該当する候補群を取得
2. 候補を生成ルールで評価
3. 複数の上位候補を略語候補として出力

この手順は，言語学の最適性理論[5]を参考にしたものである．子の理論は音韻論を対象としたものであるが，複雑な規則を用いて表層的な構造の変化を説明する考え方は，本稿の略語の生成にも適用できると判断した．

この手順２については，２節で述べた３つのルールを用いる．次節で，その評価方法について述べる

## 3.2　ルールと評価方法

評価方法は，図２に示すように，$n$ 個の略語候補 $a_1 \cdots a_n$ のそれぞれに対して，生成ルールごとに評価値を算出して，評価値の合計の大きい順に出力する．

| 略語候補 | 形態素 | 文字数 | 読み | 評価値 |
|---|---|---|---|---|
| $a_1$ | $m_1$ | $n_1$ | $y_1$ | $m_{1+}\ n_{1+}\ y_1$ |
| : | : | : | : | : |
| $a_n$ | $m_n$ | $n_n$ | $y_n$ | $m_{n+}\ n_{n+}\ y_n$ |

図２．評価方法

**(1)　形態素の省略**

このルールでは，２節でも述べたように，４通りの法則が存在している．しかし，各形態素に対して，どの法則を適用すべきかを判断することは難しい．そこで，訓練データから形態素ごとの法則の適用頻度を調べ，その情報を形態素辞書に付加した．

生成された全ての略語候補に対して，原語の形態素のそれぞれに，どの法則が適用されればその略語候補が生成されるかを考え，辞書に付加した適用頻度の値を使用して，ある略語候補 $a_k$ に対する評価値 $m_k$ を算出する．その式は以下のようになる．

$$m_k = P_{\mathrm{m}}(a_k) = \sum_{1}^{\text{形態素数}} \frac{(\text{法則の適用頻度})}{(\text{形態素の出現頻度})}$$

例としては「高等学校」から「高校」という略語候補が得られたとする．この略語候補は「高等・学校」の２つの形態素に対して，それぞれ「先頭文字・後方文字」の法則が適用された場合に生成される．そして，「高等」の先頭文字の法則，「学校」の後方文字の法則，それぞれの法則の頻度と元の形態素の頻度から，上記の式で計算する．

**(2)　文字数の減少**

略語の文字数は２文字または３文字になりやすいが，それ以上になる場合も考慮する必要がある．訓練データから，原語の文字数と略語の文字数の関係を調査して，生成されやすい文字数の略語候補の評価値が高くなるようにした．評価値 $n_k$ を求める式は以下のようになる．

$$n_k = P_n(a_k) = \frac{(a_k\text{と同じ文字数の略語の出現頻度})}{(\text{原語と同じ文字数の原語の出現頻度})}$$

**(3)　音訓の組合せと読み**

このルールでは，モーラ数と音訓の２つのルールを合わせて扱う．２章で述べたように，語が省略される場合は１フット，つまり２モーラ以上の長さを持つ．最小が２モーラであることから，略語の読みは２＋２モーラになりやすい．また，モーラ数や重箱・湯桶読み回避のために音訓が変化する場合がある．漢字の音訓情報の取得には SKK 辞書 (http://openlab.jp/skk/wiki/wiki.cgi)を使用した．

この辞書を用いて略語候補に複数の読みを付加し，モーラ数や音訓の組み合わせを考慮して，略語候補に読みの評価値 $y_k$ を以下の式で算出する．

$$y_k = P_y(a_k) = (\text{モーラ数の評価値}) + (\text{音訓の組合せの評価値})$$

# 4　実験と考察

## 4.1　実験

本稿での音訓のルールの有効性を検証するために，音訓の組合せと読みのルールを適用した場合としなかった場

表１．実験結果

| | 音訓評価なし | | 音訓評価あり | |
|---|---|---|---|---|
| | 再現率 | 精度 | 再現率 | 精度 |
| 上位1位 | 0.493 | 0.493 | 0.514 | 0.514 |
| 上位2位 | 0.635 | 0.327 | 0.622 | 0.311 |
| 上位3位 | 0.676 | 0.225 | 0.689 | 0.230 |
| 上位4位 | 0.736 | 0.184 | 0.730 | 0.183 |
| 上位5位 | 0.791 | 0.158 | 0.770 | 0.154 |
| 上位10位 | 0.845 | 0.085 | 0.872 | 0.087 |
| 平均順位 | 28.03 | | 27.15 | |

合とで実験をおこなった．実験には，Web 上のフリー百科事典 Wikipedia(http://ja.wikipedia.org/wiki/)の「漢字略語一覧」の項目に掲載されている原語・略語のセットの中から，本稿における略語の定義に一致し，訓練データと重複しないもの 148 組を用いた．これらの原語に対して，ChaSen(http://chasen-legacy.sourceforge.jp/) を用いて形態素解析をおこなったが，適切な結果が得られなかったので，想定どおりにおこなわれたと仮定して，事前に人手で形態素ごとに分かち書きした．

ここで，以下のように再現率と精度を定義する．

$$\text{再現率}_n = \frac{(\text{上位}n\text{位のうち，本来の略語と一致する数})}{(\text{データ数})}$$

$$\text{精度}_n = \frac{(\text{上位}n\text{位のうち，本来の略語と一致する数})}{(n \times \text{データ数})}$$

それぞれ，再現率は，上位 $n$ 位で本来の略語が得られる確率，精度は，本来の略語を得るためのコストを示している．再現率と精度に加え，上位何位までを出力すればよいかの目安として，検証データ全体での本来の略語の一致した平均順位も取得した．実験結果を表１に示す．表の左側は，３つのルールのうち，音訓を含む読みのルールを適用しなかった場合の結果で，右側は３つのルール全てを適用した場合の結果である．

なお，略語候補を「各形態素の先頭文字を組み合わせ」という最も単純なものにした場合，その再現率は 0.243 となった．

### 4.2 考察

２つの実験結果を比較した．上位１位の再現率が 0.493 から 0.514 へ，上位 10 位の再現率が 0.845 から 0.872 へと，それぞれ改善された．また，平均順位も 28.03 位から 27.15 位へと改善された．他に，一部に順位を下げた語句があるが，全体としては改善された語句の方が多かった．

各形態素の先頭文字の組み合わせを略語候補とした場合の 24%と比較すると，その結果の差は明らかで，上位５位までで 77%，上位 10 位までで 87%の正解の略語を得ることができた．結果の例を表２に示す．

| 慶應義塾大学：慶大 | | |
|---|---|---|
| | 音訓評価なし | 音訓評価あり |
| 上位１位 | 慶義 | **慶大** |
| 上位２位 | 慶義大 | 慶義大 |
| 上位３位 | **慶大** | 慶義 |
| 上位４位 | 義大 | 慶塾 |
| 上位５位 | 慶應義大 | 義大 |

図２．出力結果の例

## ５　結論

本稿では，複数の略語生成ルールを用いた略語自動生成について述べた．提案手法によって，より確かな略語候補を複数取得することができる．本稿の要点である音訓のルールについては，若干ではあるが有効性が確認できた．

今後の課題として，今回のシステムの性能をさらに高めることを考えている．そのためには，音訓だけではなく音韻関係を含めて，さらに読みの省略ルールを強化する，別の生成ルールとして同音異義語と比較する，などが考えられる．

## 参考文献


[1] 桜井裕，佐藤理史：ワールドワイドウェブを利用した用語説明の自動生成，情報処理学会論文誌，Vol.43，No.５，pp.1470-1480,2002.

[2] 村山紀文，奥村学：Noisy-channel model を用いた略語自動推定，言語処理学会 第 12 回年次大会，pp.763-766，2006.

[3] 石野博史：マスコミによく出る短縮語・略語解読辞典，創拓社，1992.

[4] 田窪行則，前川喜久雄，窪薗晴夫，本多清志，白井克彦，中川聖一：岩波講座 言語の科学２ 音声，岩波書店，1998.

[5] 田中章：最適性理論と日本語のいくつかの問題．新潟経営大学紀要 vol.3,pp.191-208,1997.