E-005

共起頻度を用いた略語の自動推定

Automatic Estimation for Abbreviated Words Using Co-Occurrence Rate

三輪 貴大† 大工廻 史裕† 浦谷 則好†
Takahiro Miwa Fumihiro Takue Noriyoshi Uratani

1. はじめに

「略語」とは，ある言葉（以降，原語とする）を省略や簡略化することで作り出される同義語である．本稿では，略語を「原語に存在する文字を原語の文字順序で用いて，原語の一部を省略して構築された簡略・短縮語」とする．

原語と略語の関係を把握することは，文章要約などの字数制限のある文章作成，Web 検索やテキストマイニングなどにおける網羅性の確保のために有用である．しかし，略語は一般の人々によって作成され流布されるため，他の同義語とは異なり流動的であり，未知語として処理されてしまうこともしばしばある．

これまで略語の獲得や推定に関しては，様々な手法が試されてきた[1][2][3][4]が，原語の形態素数に着目して分類し，原語の形態素数によって手法を変化させているものはない．原語の形態素数ごとの特性を詳しく把握できれば，原語の形態素数に対応させて略語候補を減少させることや，略語推定に寄与できると考えられる．

そこで，本研究では原語の形態素数ごとの略語構築における特徴を考慮し，共起頻度を用いて，表層的情報から自動的に略語を推定する手法を提案する．2 形態素と 3 形態素で構成される原語について略語推定手法を考案したので，それぞれの実験結果について報告する．

2. 既存研究

2.1 漢字構成語を対象とした研究

岡田らの研究[2]は，漢字で構成された原語から略語を自動推定するものである．この研究は，漢字の音訓情報や略語の読みを考慮するところに特徴がある．

漢字構成の＜原語，略語＞対から略語生成ルールを取得し，そのルールを用いて生成した略語候補群から最適なものを出力する略語推定システムである．訓練データとして，文献[5]から取得した漢字構成複合語（原語）とその略語の対 678 組を用いている．ルールは「形態素の省略」，「文字数の減少」，「音韻の組み合わせと読み」の 3 つである．

形態素の省略では原語を形態素に分割して，その変化法則を考える．特に，3 文字で構成される形態素では，中間文字を抜かした省略は確認できなかったと報告している．手法としては，形態素の省略を連結して取得した略語候補に対して，3 つのルールによる評価を算出し，その総和を略語候補の評価として推定する．

†東京工芸大学大学院工学研究科電子情報工学専攻
uratani@cs.t-kougei.ac.jp

2.2 カタカナ語を対象とした研究

桝井らの研究[3]は，カタカナ語で構成された原語から略語を自動生成するものである．原語から略語を生成するために略語成立のタイプを定め，原語を WWW 文書で検索し，カタカナ語の中で，生成タイプに一致した語を抽出し，WWW での出現数や接辞の一致から略語であるか否かを推定する．

WWW 文章を利用したカタカナ省略語生成システムは，以下に示す 3 つのモジュールから構成される．

- WWW 文書取得部
  原語を Web 検索し，検索結果の上位に現れる WWW 文書を取得し，前処理をする．
- 略語候補取得部
  WWW 文章から，原語と同じ字種構成のものを抽出し，その中から略語成立タイプに合致するものを略語候補として取得する．
- 略語候補判定部
  略語候補の分類によって判定を行う．しかし，それでも判定ができなければ，原語と同じ接辞を取るか否か，また高い共起頻度を持つか否かを調べ，略語かどうかを判断する．

2.3 日本語を対象とした略語推定

和田らの研究[4]は CRF（条件付き確率場）の素性にモーラ・シラブルを用いた略語の自動推定である．略語とは人間によって作られるので人間の感覚というものが大きく影響していて，音韻論の観点から略語にモーラが深いかかわりを持つという研究結果を基にしている．

モーラとシラブルを CRF の素性に用いて，日本語を対象とした略語の自動推定を行うため，MeCab を用いて原語を区切っている．その結果，モーラが略語において重要な役割をもつことを確認している．

手順としては原語xをx_1からx_nに区切り，それぞれの省略形y_1からy_nを略語ラベルとして付与する．このとき可能性として考えられる略語ラベルを連結したものを略語候補zとし，複数生成する．この略語候補zに対してCRFを用いて評価し，もっともらしい略語を選択する．CRFの素性としては原語・略語の特徴を表すものとして次のものを用いている．

観測素性f_o：原語・略語間の特徴を表す．ある原語x_iに対してどのような略語ラベルy_iが付加されるかを示す．

遷移素性f_t：略語間のモーラ・シラブルのつながりを表す．略語ラベルy_iの持つモーラ数に続いて，y_{i+1}が何モーラであるかを示す．

この方法だけでは，推定精度が 10%程度になってしまうことから推定略語の絞り込みをいくつか行い，それぞれを比較している．

その方法は略語長による絞り込み，共起頻度による絞り込みの二つに大別できる．共起頻度では，Jaccard 係数と Simpson 係数の二通りを試している．

3. 研究概要

3.1 略語の分類

Wikipedia[6]，「マスコミによく出る短縮語・略語辞典」[5]，「新聞によく出るカタカナ語・略語ハンドブック」[7]などから収集した＜原語，略語＞対を分類した．

本項では原語の形態素数による分類と，略語形成パターンによる分類について示す．

3.1.1 原語の形態素数による分類

保有している原語の重複を許した＜原語，略語＞対 942 組について，原語を構成する形態素の数に基づいて行った分類を以下の表 1 に示す．

表 1　原語の形態素数による分布

構成形態素数	＜原語，略語＞対の数（組）	割合（%）
2 形態素	300	31.85
3 形態素	275	29.19
4 形態素	150	15.92
5 形態素	86	9.13
6 形態素	69	7.32
7 形態素	38	4.03
8 形態素	20	2.12
9 形態素	3	0.32
10 形態素	1	0.11

3.1.2 略語形成パターンによる分類

岡田ら[2]，桝井ら[3]の研究でも行われていたように，語の省略部位による分類を形態素ごとに適応させ，その連結結果を略語形成パターン（以降，パターンとする）とし，保有している＜原語，略語＞対に付与した．

形態素ごとに省略形を付与し，その連結でパターンを構築する．例えば，2 形態素構成原語「修士論文」における第 1 形態素「修士」においては表 2 のようになる．

表 2　省略形の生成例

原語	省略形	省略結果
修士	後略（F）	修
	前略（B）	士
	不略（A）	修士
	総略（N）	—

これを基本として最終形態素まで行い，連結させると略語候補のパターンを得ることができる．

3 文字構成形態素の場合は先頭＋後尾という省略形を考慮しない[1][2]ため，表 2 に前後略（I），先頭字＋中字（Merge1），中字＋後尾字（Merge2）の 3 つが加わる．

本研究では接辞を 1 形態素として扱うので，4 文字以上で構成された形態素は確認できなかった．そのため考慮しないものとした．

「修士論文」の略語候補を例に取ると，「修論」は FF 型，「修士」は AN 型となり全部で 16 パターンが上げられる．

2 形態素で構成される原語において，＜原語，略語＞対 300 組みでは表 3 のような割合となった．同様に 3 形態素で構成された原語において，＜原語，略語＞対 275 組みでは表 4 のような割合になった．

表 3　2 形態素構成の原語における略語のパターン分布

パターン	略語数（件）	割合（%）
FF	205	68.33
AN	30	10.00
AF	18	6.00
BF	15	5.00
FB	14	4.67
FA	8	2.67
AB	4	1.33
NA	3	1.00
BB	2	0.67
FN	1	0.33

表 4　3 形態素構成の原語における略語のパターン分布

パターン	略語数（件）	割合（%）
FFN	75	27.27
FFA	48	17.45
FFF	27	9.82
NFF	21	7.64
FNF	16	5.82
AFN	14	5.09
ANN	12	4.36
ANF	11	4.00
AFF	8	2.91
NAA	7	2.55
BFN	6	2.18
ANA	6	2.18
FAN	4	1.45
FBN	3	1.09
AFA	3	1.09
BNA	2	0.73
AAN	2	0.73
FBA	1	0.36
FAF	1	0.36
FNB	1	0.36
BFA	1	0.36
BBN	1	0.36
BAN	1	0.36
BNB	1	0.36
AFB	1	0.36
NFB	1	0.36
NAN	1	0.36

3.2 システムの構成

本研究で構築したシステムは以下の図 1 に示す構成を取っている．

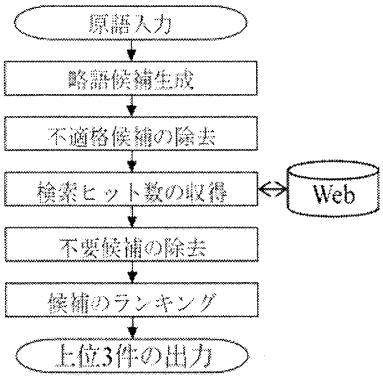

図 1　システムの構成

3.3 略語候補の生成方法

本研究では，岡田らの研究[2]を発展させた形で，略語候補を構成している．略語は原語を構成する各形態素の省略語から構成されるものとしている．つまり，原語が「修士論文」の場合，略語は第 1 形態素の「修士」と第 2 形態素「論文」に対する省略語の連の中にある．

つまり，3.1 の表 2 で示したように 2 文字構成形態素の場合は，4 つの省略形を考えれば良い．また，1 文字構成形態素は文字が有るか無いかの 2 つの省略形のみを，3 文字構成形態素は 3.1.2 文中で示したように 2 文字構成より 3 つ増加した 7 つの省略形を考慮すれば良い．

したがって，原語を構成する 1 文字，2 文字，3 文字構成形態素の個数をそれぞれx，y，zとすると，略語候補の生成総数（C_{max}）は以下の 式（1）のようになる．

$$C_{max} = 2^{x+2y} \cdot 7^{z} \qquad (1)$$

ただし「大学」，「銀行」が原語の構成形態素として存在した場合には特例として，自身の形態素の省略形を F，A，N の 3 パターンのみとした．

3.4 不適格候補の除去

3.3 で生成される略語候補総数には不適格と考えられる候補も含まれる．

ここでいう不適格とは，原語そのもの，全形態素が総略になったもの，1 文字で構成されるものである．これらは 2 形態素構成原語の略語候補群について考えると，AA，NN，FN，BN，NF，NB の 6 パターンとなる． また，3 文字構成形態素を考慮に入れれば，IN や NI も略語としては正しくない．

つまり，1 形態素を構成する文字数をベースに考えずとも A の連続，N の連続の 2 つは考慮に値しない．また，1 形態素を構成する文字数をベースに考えれば以下の場合は考慮に値しないことになる．

1 文字構成形態素では自分の省略形が A であり，前後の全ての省略形が N である場合の 1 パターン．2 文字構成形態素では自分の省略形が F，B のどちらかであり，前後の全て省略形が N である場合の 2 パターン．3 文字構成形態素では，自分の略語が F，B，I のどれかであり，前後の全ての省略形が N である 3 パターン．

そのため，推定で用いるべき略語候補数（C）は以下の式（2）のようになる．

$$C = C_{max} - (2 + x + 2y + 3z) \qquad (2)$$

3.5 不要略語候補の除去

収集して分類した＜原語，略語＞対の傾向から 2 形態素，3 形態素の略語として出現しなかったパターン（未出現パターン）について考える．

2 形態素，3 形態素どちらにおいても I を用いたパターンは出現しなかった．

また，3 形態素においては，FFB，FBF，FBB，FAB，FAA，FNF，FNA，BFF，BFB，BBF，BBB，BBA，BAF，BAB，BAA，AAF，AAB，ANB，NFA，NBF，NBB，NBA，NAF，NAB，NNA が出現しなかった．

このため，I が含まれるパターンと上に列記したパターンについては，本研究では考慮しないものとする．

また，AN，NA，ANN，NAN などの 1 形態素抜き出しパターンは，原語における部分文字列であり，それだけで単語として成立してしまうパターンである．略語として一定以上の 出現数があったとしても，このパターンが全体としての略語推定精度の向上を阻害している可能性は大いに有る．

また，3 形態素では AAN と NAA が 1 形態素抜き出しではないが部分文字列であり，推定精度の向上を阻害している可能性は高い．これら 6 パターンは，ランキングがあまりにも上位で，かつ実際の略語として出現する確率の低いパターンとなっているか否かを調査する．

3.6 共起頻度を用いた略語の推定

本研究では共起頻度の算出のため，原語と略語に関する Web 検索ヒット数を取得する必要がある．そのため，Yahoo!Japan の提供する Web 検索 API[9]を用いた．

検索するのは，以下の 3 つである．鉤括弧で括った部分は実際に Web 検索で使用するクエリである．－記号は NOT，スペース記号は AND を表す．

i. 原語と略語を同時に含む検索ヒット数（α）
「"原語" "略語"」
ii. 略語候補を含まない原語だけの検索ヒット数（β）
「"原語" -"略語"」
iii. 原語を含まない略語だけの検索ヒット数（γ）
「"略語" -"原語"」

共起頻度としては Jaccard 係数を用いて，略語の推定を行う．Jaccard 係数は式（3）に示すように，2 つの集合 X と Y において，共通要素数を少なくとも一方にある要素の総数で割ることで算出できる．

略語と原語について上記の α，β，γ を用いて式（3）に適用すると式（4）のようになる．

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \qquad (3)$$

$$= \frac{\alpha}{\alpha + \beta + \gamma} \qquad (4)$$

式（4）で求めた係数値によって順位付けを行い，その結果を推定結果とした．

4. 実験と考察

4.1 実験

以下に本研究で提案した共起頻度を用いた略語の自動推定における実験結果を示す．

表 5 は原語の重複を許さない 2 形態素構成原語 292 語について，表 6 は原語の重複を許さない 3 形態素構成原語 260 語について提案手法を行った結果を示す．3 形態素においては，ANA パターンは正答略語数は 1 つだが，上位に来る確率が高いため，無視した場合についても実験した．

各表の Base Line は 2 形態素，3 形態素構成原語に対して，3.1 の表 3 と表 4 から得られた正答略語数が多いパターンから順に正答数を割り振ったものである．

最高順位の正答略語のみ数えることとし，原語の各順位における正答率（R）は式（5）で定義する．

$$R = \frac{\text{正答略語数}}{\text{原語総数}} \tag{5}$$

表 5　2 形態素構成原語における結果

	1 位	2 位	3 位
Base Line	205 件 (68.33%)	30 件 (10.00%)	18 件 (6.00%)
追加なし	146 件 (50.00%)	86 件 (29.45%)	35 件 (11.99%)
NA なし	184 件 (63.01%)	77 件 (26.37%)	19 件 (6.51%)
AN なし	155 件 (53.08%)	76 件 (26.03%)	18 件 (6.16%)
AN・NA なし	213 件 (72.95%)	32 件 (10.96%)	9 件 (3.08)

表 6　3 形態素構成原語における結果

	1 位	2 位	3 位
Base Line	75 件 (27.27%)	48 件 (17.45%)	27 件 (9.82%)
追加なし	5 件 (1.92%)	48 件 (18.46%)	82 件 (31.54%)
NAN なし	5 件 (1.92%)	48 件 (18.46%)	100 件 (38.46%)
ANN なし	5 件 (1.92%)	44 件 (16.92%)	94 件 (36.15%)
AAN なし	39 件 (15.00%)	91 件 (35.00%)	65 件 (25.00%)
NAA なし	15 件 (5.77%)	109 件 (41.92%)	65 件 (25.00%)
AAN・NAA なし	113 件 (43.46%)	71 件 (27.31%)	30 件 (11.54%)
NAA・AAN・ANN なし	122 件 (46.92%)	62 件 (23.85%)	32 件 (12.31%)
NAA・AAN・NAN なし	130 件 (50.00%)	76 件 (29.23%)	22 件 (8.46%)
4 パターン全なし	140 件 (53.85%)	66 件 (25.38%)	16 件 (6.15%)
4 パターン・ANA なし	152 件 (58.46%)	60 件 (23.08%)	12 件 (4.62%)

4.2 考察

2 形態素構成原語における NA，AN を無視する方法の単独活用は，一定以上の有効性を見せている．特に，両方を同時適用した推定では著しい精度の向上が図れた．

3 形態素構成原語における NAN，ANN を無視する方法の単独活用は何もしないものに比べて 3 位の正答率を向上させることには成功しているが，ANN に関しては 2 位が減少するマイナス面も現われている．また，AAN，NAA を無視するパターンの単独活用は 1，2 位に対して正答率を向上させる結果を得た．複数同時活用では，著しい正答率向上が見られた．これは，各方法が別々の原語に関して正答順位の向上に寄与したと考えられる．

最後に，3.5 で触れた省略形 I に関しては保有している＜原語，略語＞対で，全く出現していないことが判明した．このことから，このパターンを用いる略語を最初から除外して考えることで推定の手間を少なくできると考えられる．

5. 結論

原語の構成形態素数ごとの特徴による略語候補の生成手法は，略語候補数の削減において有効性を示した．略語候補数の削減は略語推定において精度に大きく寄与することが確認できた．

また上位 3 位における精度は，2 形態素の Base Line では 84.33%，本手法の最優結果では 86.99%と僅かな向上となったが，1 位正答率に関しては 4%強の向上が得られた．3 形態素では Base Line では 54.54%，本手法の最優結果では 86.16%と大幅な向上が得られた．結果として，本提案手法は略語推定にいて高い優位性を示せた．

今後は，手法を 4 形態素以上へ拡大したい．さらに，カタカナ字種への対応法も開発していきたい．

参考文献

[1]大工廻史裕，三輪貴大，浦谷則好：“確立モデルを用いた略語の自動推定”，FIT2009（第 8 回情報科学技術フォーラム）講演論文集，E-024，pp.317-318，2009.

[2]岡田真，高橋幹浩：“漢字を中心とした複合語の略語の自動生成”，言語処理学会第 14 回年次大会発表論文集，C4-4，pp.787-789，2008.

[3]桝井文人，松田良一，野呂康洋，河合敦夫，井須尚紀：“World Wide Web を知識源としたカタカナ省略語の自動生成”，2004 年度電磁情報通信学会基礎・境界ソサイエティ大会講演論文集，A-13-1，pp.527-528.

[4]和田健太，近山隆：“素性にモーラとシラブルを用いた略語の自動推定”，情報処理学会全国大会講演論文集，1X-1，pp.2，2010.

[5]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.

[6]Wikipedia：http://ja.wikipedia.org/wiki/

[7] 大薗友和：“新聞によく出るカタカナ語・略語ハンドブック”，日本実業出版社，1989.

[8]Yahoo! Japan API：http://developer.yahoo.co.jp/