E-051

共起頻度と略語形成パターンを用いた略語の自動推定

Automatic Estimation for Abbreviated Using Co-Occurrence Rate and Composite Patterns

三輪 貴大† 浦谷 則好 †
Takahiro Miwa Noriyoshi Uratani

1．はじめに

「略語」とは，ある言語（以降，原語）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言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．すでに著者らは略語の自動推定手法を提案し，実験結果を報告している[1]．しかし成果を詳細に分析したところ，取得した検索ヒット数に問題があることがわかった．そこで本稿では，まず前年度研究の検証を行い，検証によって修正された検索ヒット数に基づいて進めた本研究の実験結果を報告する．

提案手法は，2 形態素原語の略語構築における特徴を考慮し，原語，略語の頻度及びその両者の共起頻度を用いて，表層的情報から自動的に略語を推定する手法である．

2．前年度研究の検証

前年度発表した研究における問題点を検証し，改善を図る．以降では原語の検索ヒット数を O，略語候補の検索ヒット数を A で表す．

2.1 前年度研究の手法概説

システムの構成を図 1 に示す．

略語を形成するために各形態素の一部が省略されるが，省略部位による分類を省略形と呼ぶ．省略形の連結により略語形成パターン（以降，パターン）が決定される．表 1 に省略形を，表 2 にパターンの例を示す．

推定手法の概説は次の通りである．パターンを網羅することで略語候補を生成し，略語として不適格である 1 文字パターンや AA，NN パターンを削除する．不適格ではない候補について，Web における検索ヒット数を取得する．そして，収集した＜原語，略語＞対データベースにおいて出現しないパターン（I 含有，BA）を不要候補として削除（以降，特定パターン削除）する．最後に，不要候補でないものは Jaccard 係数を算出し，ランキングする．このランキング結果の上位 3 件を推定結果とする．

このとき部分文字列であり，かつ正当な略語を持つパターン（AN，NA）のいずれか，または両方を不要候補に含めた場合についても実験を行うものとする．

研究成果として，I 含有パターンは 2，3 形態素の正当略語とならないことが，収集した＜原語，略語＞対データベースで確認された．

†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

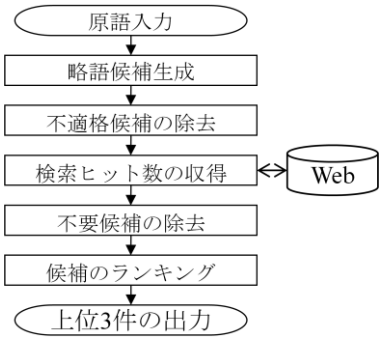

図 1　システムの構成

表 1　省略形一例

原語	省略形	省略結果
北海道	後部略（F）	北
	後尾字略（H）	北海
	前後略（I）	海
	前部略（B）	道
	先頭字略（R）	海道
	不略（A）	北海道
	総略（N）	—

表 2　略語形成パターン一例

原語	パターン	省略結果
修士　論文	FF	修論
	……	……
	NA	論文
	NN	—

2.2 検索ヒット数の取得における問題

Yahoo! API を用いた検索ヒット数の取得において，取得された値が 0 になってしまうエラーが確認できている．

また，O と A の AND 検索結果を $0 \cap A$ とした下記の式(1)が満たされてしまうエラーも確認されている．

$$(0 \cap A) > Min(0,\ A) \qquad (1)$$

この現象は，Yahoo! API，手動による Yahoo!，Bing API，手動による Bing や Google の全てで確認できた．特に，略語候補が原語の部分文字列の場合に多い．

比較的 Google 検索エンジンでは，この影響が小さいことが確認できた．このとき問題が確認されたのは AB，BB，FA，FB，FF，FH，HA と部分文字列パターンであった．全有効データは 3154 件，エラーデータは 540 件

（17.12%）であった．このうち，エラーデータにおける非部分文字列は 115 件（21.30%），部分文字列は 425 件（78.70%）となった．

2.3 語の順序による検索ヒット数の変動問題

Yahoo!，Bing では AND 検索において語の順序によって検索ヒット数が大幅に変化する場合が度々見られた．

本研究では，語の順序によって検索ヒット数が変わってしまうと，どちらを推定に用いるべきかの判断がつかないため，この問題は極力排除したい．

Google における Web 検索においては，検証中にこの問題が確認されることはなかった．

2.4 前年度研究の検証に対する結論

2.1 で示した手法による前年度研究は，2.2，2.3 で述べたような問題が存在している．そのため，この研究成果は信頼性が著しく低い．だが，これらの問題は手法自体の問題点ではなく，検索ヒット数を Web から取得することに関しての問題点である．よって，検索ヒット数の取得，そして取得した数値に対する修正を行うことで手法自体の有効性を検証できると考える．

3. Jaccard 係数の補正による推定手法の改良

3.1 Google から取得した検索ヒット数の修正

前述の理由より，使用する Web 検索エンジンは Google とする．また，取得してデータベース化した検索ヒット数に対して 2.2 に示した式(1)が満たされてしまうものは以下に示す式(2)で$O \cap A$ を補正することで論理的整合性を与えることとする．

$$O \cap A = \mathrm{Min}(O,\ A) \qquad (2)$$

3.2 入力原語及びその略語の整理

前年度研究において入力に用いていた原語，全 292 件について整理を行った．

以下に示す例のような略語が原語の意味を保持するためにはある特定のコミュニティや話題を必要とするものがある．このような略語が多義語になるものに関しては，一般的な原語と略語の関係であっても削除した．略語を複数もつ原語では，該当する略語のみを削除することとした．

- 定期預金 → 定期
- 明治大学 → 明治
- 宮崎交通 → 宮交
- 宮城交通 → 宮交

この入力原語の整理の他にも，パターンの特定間違い，原語の形態素区分箇所の間違いなども修正した．これらの整理・修正処理の結果，入力に用いる原語数は 254 件に減少した．この整理後の略語のパターン出現数と割合を表 3 に示す．

表 3 パターンの出現数と割合

パターン	出現数（件）	割合（%）
FF	170	65.64
AN	24	9.27
AF	19	7.34
BF	16	6.18
FB	14	5.41
FA	10	3.86
NA	2	0.77
BB	2	0.77
HN	1	0.39
AB	1	0.39

3.3 修正済み検索ヒット数を用いた推定結果

前年度の研究手法に対して，2.2，2.3 で述べた問題をできるだけ回避するために検索ヒット数の修正（3.1）と不適切な略語の削除（3.2）を行い，略語の自動推定を行った結果について表 4 に示す．

ベースラインとは保有略語データベースから出現数の多いパターン順に取得した場合のものである．また，前年度手法の結果においては不要候補に AN，NA を入れない場合（ノーマル）と，不要候補に AN，NA のいずれか，または両方を含めた場合（パターン＋削除）における最も精度の高い結果を掲載した．

前年度の手法において有効だとした 2 形態素における AN，NA パターンを削除する方法は，全 3 パターンの全てで正当率の低下を起こしている．これは AN，NA パターンを削除しても正当が上昇せず，かつすでに上位に存在した正当の AN，NA パターンが除外されてしまうために起こると想定できる．そこで，表 5 に結果の 4 位までに出現したパターンとその数を正当件数／総数として示す．

表 5 に示されている，BH，BA，RN，AH，AB，AR，NH，NR の 8 パターンは，収集したデータベースでは正当な略語が保有されていないパターンである．この中でも BA は入力総数が 254 件であるが，その内 157 件で 1～4 位に含まれる．

そこで，推定において正当略語を邪魔するようなパターンの Jaccard 係数に対する補正処理を試みる必要がある．

また，前述した正当略語を持たない 8 パターンのうち，AB と AR を抜く 6 パターンが部分文字列である．さらに，部分文字列でない AB と AR のパターンはともに上位に出現しにくいことが表 5 から見て取れる．このことから，部分文字列であるパターンが推定に対して悪影響を与えていると考えられる．

表 4 ベースラインと前年度手法の実験結果

	ベースライン	前年度手法	
		ノーマル	NA 削除
1 位正当数 (正当率)	170 件 (65.64 %)	59 件 (23.23 %)	57 件 (22.44 %)
1～3 位 正当数 (正当率)	213 件 (82.24 %)	153 件 (60.24 %)	151 件 (59.45 %)

表 5　結果の上位に現れるパターンとその数

パターン	1 位	2 位	3 位	4 位
FF	19／20	24／24	35／37	48／49
FB	1／1	2／2	3／5	4／5
FA	5／5	1／1	2／3	1／6
HN	0／2	1／2	0／2	0／2
BF	8／81	2／53	5／43	0／28
BH	0／1	0／0	0／0	0／1
BB	0／2	1／1	0／0	0／3
BA	0／44	0／42	0／35	0／36
RN	0／2	0／5	0／2	0／4
AF	15／49	4／48	0／53	0／39
AH	0／1	0／0	0／1	0／1
AB	0／0	0／1	0／6	0／3
AN	9／24	10／35	4／37	1／32
AR	0／0	0／0	0／0	0／1
NH	0／0	0／0	0／1	0／0
NR	0／1	0／0	0／1	0／0
NA	2／22	0／39	0／28	0／31

3.4 部分文字列の悪影響に対する検証

まず，2 形態素における部分文字列と成り得るパターンの全てを分類して表 6 に示す．このとき，1～4 位とは，表 5 における順位のことを指す．

ここで，略語として不適格である部分文字列は排除する．

正当な略語をもつ部分文字列においても，HN と NA の両パターンは正当数が極めて少ない．特に NA は大量の不正解候補が 1～4 位に入っている．また BF，AF，AN も大量の不正解が 1～4 位に入ってくる．

部分文字列の全データ数が 1331 件において A が O 以上になるデータ数が 1094 件（89.19%）となる．つまり，2.2 で示した式(1)を満たしてしまう部分文字列では，$O \cap A$ の補正結果は下記の式(3)で表される．

$$O \cap A = O \tag{3}$$

Jaccard 係数の算出式を Jac とすると O，A，$O \cap A$を用いて表した式(4)は式(3)を適用することにより，式(5)のように変形される．

$$\mathrm{Jac} = \frac{O \cap A}{O + A - (O \cap A)} \tag{4}$$

$$\mathrm{Jac} = \frac{O}{A} \tag{5}$$

式(5)である場合，O と A が同数に近ければ Jaccard 係数は 1 に近づく．また，A が著しく O より大きいと Jaccard 係数は 0 に近づく．

「略語として機能しない部分文字列（以降，非略語部分文字列）」について考えると、論理的には O と A は同値に近づくので，比は 1 に近づく．さらに，「略語として機能する部分文字列（以降，略語部分文字列）」について考えると，論理的には O に対して A が大きくなるので，比は 0 に近づく．

つまり，非略語部分文字列の Jaccard 係数が 1 に近づき，略語部分文字列の Jaccard 係数が 0 に近づいてしまう．したがって，Jaccard 係数の大きい順に順位付けを行うと非略語部分文字列が高い順位に，略語部分文字列が低い順位になるという問題が起きる．

表 6　部分文字列と分類

略語として不適格	1～4 位に出現		1～4 位に未出現
	正当略語を持つ	正当略語を持たない	
FN	HN	BH	RF
BN	BF	BA	RH
IN	AF	RN	RA
NF	AN	AH	
NI	NA	NH	
NB		NR	

3.5 改良された推定システムの構成

本研究において 2.1 で示した手法を改良し構築したシステムの構成を図 2 に示す．2.1 のシステム構成との変更点は，「不要候補の除去」が「Jaccard 係数の補正」に変更されたことである．また，本手法では前年度研究で不要候補とした I 含有パターンは，不適格候補として推定対象から排除している．

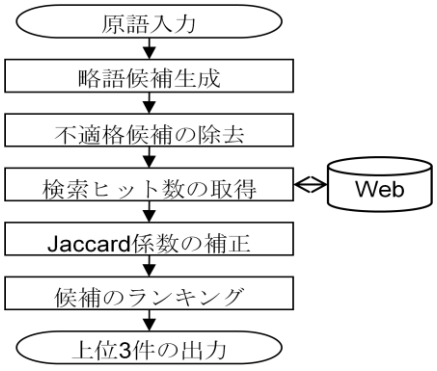

図 2　システム構成

3.6 Jaccard 係数の補正

本研究では 2.1 で示した手法とは異なり，「特定不要パターン削除」は採用せず Jaccard 係数の補正処理によって精度の向上を図る．

この補正処理では 3.5 で示した問題を解決するために，以下の式(6)を満たす部分文字列全てに対して A の補正処理を下記の式(7)によって行うこととした．

$$\frac{A}{O} > 1 \tag{6}$$

$$A = O\left(\alpha * \log_e \frac{A}{O} + 1\right) \tag{7}$$

すなわち式(7)は略語のほうが原語よりも一般的に使われている場合に，Jaccard 係数が下がることを抑えるためのものである．

言い替えると，式(7)は A の大きさを O に近づける処理であり，これによって Jaccard 係数は上昇する．式(7)における α は 0.2171 としたが，これは A/O が 1 のときに $A = O$ となり，A/O が 100 のときに $A = 2 * O$ となるように調整した結果である．α を小さくするほど，調整量は大きくなる．

4. 実験と考察

4.1 実験

以下に 3.で示した「Jaccard 係数の補正による推定手法の改良」における実験結果を表 7 に示す．式(7)による補正を log 補正と呼ぶことにする．3.4 で示したように部分文字列は Jaccard 係数が高くなる傾向がある．よって，「部分文字列に対しては一定の数値で除算する Jaccard 係数の補正（以降，除算補正）」を行う．表 7 では log 補正のみを行い，部分文字列の Jaccard 係数を除算しない場合の列名を 1 とした．また，log 補正と同時に部分文字列を除算する場合は除算に用いた倍率を列名とし，精度の高い倍率周辺を示す．さらに，部分文字列を完全に削除した実験結果についても表 7 で「削除」として示す．

また，表 8 に各正当パターンごとに本手法で取得した正当件数を総正当件数で割った割合を，その件数（本手法で得られた正当件数／総正当件数）とともに掲載する．

表 7　log 補正と除算補正による実験結果

	1	1/200	1/300	1/400	削除
1st	35	199	200	198	185
2nd	16	23	23	22	25
3th	60	14	9	12	11
4th	36	8	9	11	10
1st Rate	13.78	78.35	**78.74**	77.95	72.83
1~3 Rate	43.7	**92.91**	91.34	91.34	87.01

表 8　パターンごと 1～3 位における正当取得件数

パターン	略語数（件）	割合 (%)
BB	2／2	100.00
AB	1／1	100.00
FF	167／170	98.24
FB	13／14	92.86
FA	9／10	90.00
AF	12／19	63.16
BF	10／16	62.50
NA	1／2	50.00
AN	9／24	37.50
HN	0／1	0.00

4.2 考察

2 形態素構成原語における log 補正方法を単独で用いた場合では 1 位及び 1～3 位の推定精度に何の寄与もしない．しかし，部分文字列の除算を同時に行うことで推定精度の向上が図られた．

これは，3.6 で想定した以上に非略語部分文字列が高い順位に来ていることを示している．しかし，部分文字列を安易に削除する方法と除算する方法を比較すると，除算倍率にもよるが最大で 1 位の正当件数で 15 件（5.91%），1～3 位の正当件数は 13 件（5.90%）の差が出る．

また，本手法では HN，AN，NA パターンで正当率が低いが，HN と NA では入力原語におけるの正当数が極めて少ないためで，AN においては除算補正による影響であると考えられる．

さらに，比較対象として 3.6 の式(6)の条件に当てはまる部分文字列に対し，A を O で代替した補正方法を「A=O 補正」として行った．この結果，最高の 1 位正当率は部分文字列の除算倍率が $1/300$ ～ $1/500$ における 76.77 であった．また，最高の 1～3 位正当率は部分文字列の除算倍率が $1/200$ における 89.37 であった．しかし，A=O 補正では同値の Jaccard 係数が起こりやすい．前述した A=O 補正の正当率は，同数の Jaccard 係数では順位の重複を許さずに略語候補の生成順に順位を付与した結果である．

もし A=O 補正で順位の重複を許すならば，重複している順位において何かの基準で序列を付けなければならない．このような問題があるため，A=O 補正は補正手法として採用しにくいと考える．

5. 結論

前年度研究の手法は，「特定不要パターン削除」以外では本研究と同じ手順を用いている．このため，前年度研究の手順は 2 形態素において妥当であることがわかった．

また，本研究において行った 3.6 に示した「特定不要パターン削除」の代わりに採用した手法「Jaccard 係数の補正」による精度向上が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．

さらに 1～3 位における精度は 3.3 で示した表 4 のベースライン（82.24%）と比較して本研究の最高結果（92.91）と，1 位における精度は同様に表 4 のベースライン（65.64%）と比較して本研究の最高結果（78.74）ではどちらも向上が得られた．結果として提案手法は略語推定において高い優位性を示せた．

今後は，3 形態素以上の原語に対して手法を拡大していきたい．また，さらに他字種への対応も検討していきたい．

参考文献

[1] 三輪貴大，大工廻史裕，浦谷則好：“共起頻度を用いた略語の自動推定”，FIT2010（第 9 回情報科学技術フォーラム）講演論文集，E-005，pp.209-212，2010.

[2] 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.