モーラ数とバイグラム情報を評価基準に用いた カタカナ語の略語自動生成手法

岡田 真　中川 大輔
大阪府立大学大学院 理学系研究科 情報数理科学専攻

1. はじめに

一般に，表現の簡略化のために，名詞や名詞句(以降合わせて「原語」とする)から数文字を抜き出し，組み合わせて短縮した語(以降「略語」とする)を同義語として扱うことがある．このような略語と原語の同義語関係を把握することは，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]．このような研究としては和田らの研究[2]や村山らの研究[3]が挙げられる．

しかし，これまでの研究では，略語の自動生成についてはほとんどおこなわれてこなかった．そこで我々の研究室では，漢字で構成されている複合語に限定して略語の読みや音訓情報を用いた略語の自動生成手法を提案した[4]．

本稿では対象をカタカナで構成された複合語(以後「カタカナ語」と呼ぶ)に限定した略語の自動生成手法について提案する．本稿のシステムでは，まず既存の原語と略語の関係から略語の評価基準を取得する．そして，それらの評価基準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．今回は評価基準として略語のバイグラム情報に着目してその有効性を検証する．

本稿では，第 2 章でカタカナ略語生成システムの概要について述べ，第 3 章で略語の評価方法を詳細に説明する．第 4 章で実験を用いたシステムの有効性を検証して，考察をおこなう．最後にまとめと今後の課題について述べる．

2. カタカナ略語生成

本研究で提案するシステムは，以下の手順で生成をおこなう．

I. 原語を形態素解析する，
II. 略語生成条件に基づき略語候補群を生成する，
III. 候補を 3 つの評価基準で評価する，
IV. 評価値の高い候補から順に出力する．

原語の形態素解析については，図 2.1 に例を挙げる．また，略語生成条件は「略語は原語の各形態素それぞれ前方から取られて生成される，かつ，原語と略語の先頭文字が一致する」と定義する．略語生成条件を定めた理由はシステムの高速化を図ったためである．文献[5]で明記されているカタカナ略語 205 語のうち，原語と先頭文字が一致していない略語が 4 語，原語の各形態素それぞれの前方から取られていない略語が 14 語あった．それらの語が少ないので，今回は略語候補を生成する際，そういった略語候補を省き，略語候補数を減らすことにした．そして，略語候補を減らすことによって，システムの高速化を図った．

原語	⇒	形態素解析結果
インターハイスクール	⇒	インター/ハイ/スクール
パーソナルコンピュータ	⇒	パーソナル/コンピュータ
オリジナルカード	⇒	オリジナル/カード

図 2.1　原語と形態素解析結果の例

3. 略語評価方法

表 3.1 は，ある原語から n 個の略語候補を生成するときの評価値を示している．略語候補 a(1)…a(n)のそれぞれに対して，形態素数を用いた評価値 V_N，モーラ数を用いた評価値 V_M，バイグラムを用いた評価値 V_B と，3 つの評価値の和を得る．全ての略語候補について評価をおこなった後，3 つの評価値の和の大きい順に略語候補として出力する．以下，原語を W，その原語から生成された略語候補を R_w であらわす．

3.1. 形態素数を用いた評価値 V_N

この評価基準は，原語のそれぞれの形態素からどれだけのモーラ数がとられて略語が生成されたかを考慮するためのものである．経験的に，原語が 2 つの形態素で構成されている場合，略語は原語の第 1 形態素から 2 モーラ，第 2 形態素から 2 モーラとられて生成されている場合が多い．実際に訓練データを用いて調査した結果でも，原語の形態素数が 2 である略語 110 語中，原語の第 1 形態素から 2 モーラ，第 2 形態素か

表 3.1　略語の評価方法

略語候補	評価値
a(1)	$V_N(1)+V_M(1)+V_B(1)$
:	:
a(n)	$V_N(n)+V_M(n)+V_B(n)$

ら 2 モーラとられて生成されている略語は 70 語あった．しかしそれ以外の場合も考慮するために，訓練データ中の原語の形態素数とその略語のとられ方を調査して，生成されやすい取られ方をしている略語候補の評価値が高くなるようにした．訓練データの中で原語 W と同じ形態素数の原語の総数を N(w)，W と同じ形態素数の原語から生成される略語のうち，ある語 R_Wと同じ取られ方をしている略語の総数を$N(w, R_W)$とする．このとき評価値V_Nを求める式は以下のようになる．

$$V_N = \frac{N(w, R_W)}{N(w)}$$

同じ取られ方とは，それぞれの形態素から同じモーラ数ずつ取られていることを示す．略語ととられ方の例を表 3.2 に示す．例えば，「セリーグ」と「パリーグ」は同じ取られ方をしていて，形態素数が 2 である原語の総数が 110 語のうち，その略語が第 1 形態素から 1 モーラ，第 2 形態素から 3 モーラ取られている語は 2 語ある．よって，

$$\text{評価値} V_N = \frac{2}{110} = 0.0181818$$

となる．

3.2. モーラ数を用いた評価値 V_M

この評価基準では，原語からどれだけのモーラ数が減少して略語が生成されたかを考慮する．実際に訓練データを用いて調査した結果を表 3.3 に示す．原語のモーラ数と略語のモーラ数の交差した位置には原語略語対の数を示した．表 3.3 にもあるように，原語のモーラ数が 7 以上ならば，略語のモーラ数が 4 モーラになりやすいことがわかった．しかし，それ以外の場合も考慮するため，訓練データ中の原語のモーラ数と略語のモーラ数を調査して，生成されやすいモーラ数の略語候補の評価値が高くなるように設定した．訓練データの中で原語 W と同じモーラ数の原語の総数を M(w)，W と同じモーラ数の原語から生成される略語のうち，R_w と同じモーラ数である略語の総数を$M(w, R_W)$とするとき，評価値V_Mを求める式は以下のようになる．

$$V_M = \frac{M(w, R_W)}{M(w)}$$

表 3.2　原語と略語と形態素からのとられ方の例

原語	略語	とられ方
セントラル・リーグ	セ・リーグ	1-3
パシフィック・リーグ	パ・リーグ	1-3
カメラ・リハーサル	カメ・リハ	2-2

略語「サスプロ」の場合を考えると，原語が「サステイニングプログラム」であり，原語のモーラ数 12 から略語のモーラ数は 4 に減少している．訓練データでは，表 3.3 に示すように，原語のモーラが 12 である略語 4 語のうち，略語のモーラ数が 4 である語が 3 語なので，評価値V_Mは以下のようになる．

$$\text{評価値} V_M = \frac{3}{4} = 0.75$$

3.3. バイグラムを用いた評価値 V_B

この評価基準は，略語のバイグラムを基にしたものである．経験的に略語には「コン」などがふくまれていることが多いことが知られている（例，「パソコン」，「ミスコン」）．そういったバイグラムの略語中の出現頻度を考慮するためにこの評価値を設定する．訓練データの中で$R_W[k]$が第 1 モーラであるバイグラムの総数を$B(R_W[k])$，$R_W[k]$が第 1 モーラで $R_W[k+1]$が第 2 モーラであるバイグラムの総数を $B(R_W[k], R_W[k+1])$とする．また，ある語 A の 1 モーラ目を A[1]，2 モーラ目を A[2]，…，n モーラ目を A[n]とする．次に，A のモーラ数を mora(A)とする．よって，mora(A)-1 は A のバイグラムの数を表す．評価値V_Bを求める式は以下のようになる．

$$V_B = \frac{\sum_{k=1}^{\text{mora}(R_W)-1} \frac{B(R_W[k], R_W[k+1])}{B(R_W[k])}}{\text{mora}\,(R_W)-1}$$

表 3.3　原語と略語のモーラ数の関係

		略語のモーラ数						
		2	3	4	5	6	7	合計
原語のモーラ数	4	8	0	0	0	0	0	8
	5	18	5	1	0	0	0	24
	6	9	11	9	0	0	0	29
	7	2	17	29	0	0	0	48
	8	1	6	18	2	1	0	28
	9	4	3	21	2	4	0	34
	10	0	1	15	0	1	0	17
	11	0	2	7	1	0	0	10
	12	0	0	3	0	0	1	4
	13	0	0	1	0	0	0	1
	14	0	1	1	0	0	0	2
	合計	42	46	105	5	6	1	205

4. 実験と考察

4.1. 実験用データ

形態素数を用いた評価値 V_N とモーラ数を用いた評価値に用いた評価値 V_M の作成用の訓練データには，文献[5]に明記してあるカタカナ略語とその原語 205 語を用いた．また，バイグラムを用いた評価値に用いた訓練データには，文献[5]に明記してある全略語 698 語を用いた．

4.2. 実験

実験は，① 3 つの評価値の総和で評価したもの，② 形態素数を用いた評価値とモーラ数を用いた評価値の評価値だけで評価したもの，③ バイグラムを用いた評価値で評価したものの 3 つに分けておこなった．この目的は評価値の組み合わせをかえることでそれらの有効性を測ることである．実験データとして，Wikipedia[6]に略語，略称関係にあると明記してあるものを抜粋した結果，115 語の原語とそれらに対応したカタカナ略語が得られた．これらの原語に対して，想定どおりに形態素解析が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をおこない，それを実験に用いた．表 4.1 には本システムで生成した評価値の高い順に並べられた略語候補の中で実際の略語と一致した場合の順位と原語略語対の数を示す．かっこ内の数値は順位が同値の略語候補が複数出力された原語略語対の数である．

4.3. 考察

①と②の実験結果を比較すると，1 位の原語略語対の数が①の 51 語より②の 53 語の方が多い．

表 4.1　実験結果

順位	①	②	③
1	51(4)	53(9)	5(2)
2	7	8(1)	25(4)
3	7(1)	6(1)	7(3)
4	2	2	9(1)
5	6(1)	5(2)	9(1)
6	7	7(1)	11(6)
7	8	6	4(1)
8	4	3	4(2)
9	2	4	9(4)
10	1	2	6(3)
11 以降	7	6	13(8)
生成できず	13	13	13
計	115	115	115

バイグラムを用いた評価値を入れることにより有効性が低くなっているように見えるが，単独で 1 位になっている原語略語対の数に着目すると，①の 47 語が②の 44 語より多い．これにより，バイグラムを用いた評価値を入れることによって，曖昧性が解消されたと考えられる．しかし，③のバイグラムを用いた評価値だけで評価したものに関しては，それほどの効果を得られなかった．その原因として，バイグラムを用いた評価値を算出するために用いた訓練データ数が少ないためと考えられる．訓練データを追加して，評価の精度とバイグラムとの関連をさらに調査する必要がある．

5. まとめと今後の課題

本稿では，カタカナ語からの複数の評価基準を用いた略語自動生成手法について述べた，バイグラム情報を用いた評価については，部分的な有効性を確認した．

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，バイグラム情報に加えて，長音や促音などの特殊モーラに関しての情報を扱えるように，システムを強化することが考えられる．また，略語生成条件のため生成できなかった略語候補に対して対応できるようにシステムを強化することも課題となる．

参考文献

[1] 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, 電気情報通信学会論文誌, pp.1624-1628,2010.

[2] 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, 情報処理学会 第 190 回自然言語処理研究会, pp.67-72,2009.

[3] 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.

[4] 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−音訓を考慮したルールを用いて−, 言語処理学会 第 14 回年次大会, pp.787-789,2008.

[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/