論　文

λ 適 合 分 割*

佐　藤　睦** 北 橋 忠 宏*** 田 中 幸 吉***

Abstract

We introduced the derivative of the partition R on a vocabualy and the mached partition of the language in the paper[1]. But the theory given can strictly apply no actual example of the language, because an acutual example has exception and all sentence belongs to the language can't be given.

In this paper, for a pair (a_i, a_j) of two words, $F_R(a_i, a_j)$, grade of identity in the sense of a partition R, is defined. And, based on this definition, the λ-derivative of the partition R on a vocabulary and the λ-mached partition of the language is defiened and some properties of them are considered.

The algolithm of seaching for the λ-mached partition of the language for given $\lambda(0<\lambda\leq 1)$ is given, and simple examples of λ-matched partition are given.

1. はじめに

筆者らは文献 1) において最大適合分割を提案したが，本論文においては簡単な例を用いて最大適合分割を求め，その有用性を確かめ，問題点を見い出すことを主眼として述べる．

まず，文献 1) の理論の等号（文脈の分割 R の意味での一致）の部分は実例に適用することができないので，二つの単語 a_i, a_j の文脈の分割 R の意味での一致の度合 $F_R(a_i, a_j)$ を定義し，この F_R に基づいて，λ 誘導，λ 適合の概念を定義し，いくつかの定理を述べる．

つぎに，これらの諸定義，諸定理に基づいて，実例を処理するアルゴリズムを述べ，このアルゴリズムを適用する際の問題点を指摘する．

最後に，実例として英語の教科書を対象として，最大適合分割などを求め，その結果について考察する．

2. λ 誘導および λ 適合

この章では，次章のアルゴリズムを導くために必要な定義および定理を述べる．ただし，その定理と対応する定理が文献 1) にある場合などは，簡単に説明を付け，証明は省略する．なお，ここで述べない定義および記法は文献 1) に従う．

処理すべき与えられた言語は有限（有限個の文により構成されている）であるので，以下の議論では言語 L は有限であり，固定されているものとして添字 L を省略する．

与えられた言語 L を処理する場合には，文脈などが完全に一致することを基準にすることは適当でないと考えられるので，一致の度合を定義する．

〔定義 2.1〕 S を有限集合したとき，$n(S)$ で集合 S に属する要素の個数を表わす．

〔定義 2.2〕 R を語い Σ 上の同値関係（分割）とする．このとき，単語 a の文脈 $C(a)$ の中で，単語 b の文脈 $C(b)$ の中に R 同値なものが存在するものの集まりを $C_b^R(a)$ で表わす．

すなわち

$$C(a)=\{(\alpha,\beta)\mid\alpha a\beta\in L\}$$

$$C_b^R(a)=\{(\alpha,\beta)\mid(\alpha,\beta)\in C(a),(\alpha,\beta)/R\in C(b)/R\}$$

定義 2.2 よりつぎの 2 つの命題は明らかに成り立つ．

〔定理 2.1〕 $\phi\subseteq C_b^R(a)\subseteq C(a)$

ここで，ϕ は空集合である．

* The λ-Matched Partition by Mutsumi SATO (Faculty of science and Technology, Kinki University), Tadahiiro KITAHASHI and Kohkichi TANAKA (Faculty of Engineering Science, Osaka University).
** 近畿大学理工学部経営工学科
*** 大阪大学基礎工学部情報工学科

〔**定理 2.2**〕 $0 \leq n(C_b{}^R(a)) \leq n(C(a))$

ここで，$n(C_b{}^R(a))=0$ となるのは $C_b{}^R(a)=\phi$ のとき，すなわち，$C(a)/R \cap C(b)/R=\phi$ のときかつそのときに限る．また，$n(C_b{}^R(a))=n(C(a))$ となるのは $C_b{}^R(a)=C(a)$ すなわち，$C(a)/R \subseteq C(b)/R$ のときかつそのときに限る．

〔**定義 2.3**〕 2つの単語 a と b のそれぞれの文脈 $C(a)$ と $C(b)$ の R の意味での一致の度合をつぎの式で定義する．

$$F_R(a, b)=\frac{n(C_b{}^R(a))+n(C_a{}^R(b))}{n(C(a))+n(C(b))}$$

定義より，F_R は対称的である．すなわち

$$F_R(a, b)=F_R(b, a)$$

また，定理 2.2 より

〔**定理 2.3**〕 $0 \leq F_R(a, b) \leq 1$

定理 2.2 の説明より

〔**定理 2.4**〕 $C(a)/R \cap C(b)/R=\phi$ は $F_R(a, b)=0$ となる必要十分条件である．

〔**定理 2.5**〕 $C(a)/R=C(b)/R$ $(a \cdot \sim/R \cdot b)$ は $F_R(a, b)=1$ となる必要十分条件である．

これらの定理は，$F_R(a, b)$ を2つの単語 a と b のそれぞれの文脈 $C(a)$ と $C(b)$ の R の意味での一致の度合と定義する1つの根拠である．

ここで，簡単な例をあげておく．

〔例〕 $\Sigma=\{\text{this, that, is, a, pen, desk}\}$

$L=\{\text{this is a pen, that is a desk, is this a pen}\}$

とおけば，

$$C(\text{this})=\{(\epsilon, \text{ is a pen}), (\text{is, a pen})\}$$
$$C(\text{that})=\{(\epsilon, \text{ is a desk})\}$$
$$C(\text{is})=\{(\text{this, a pen}), (\text{that, a desk}), (\epsilon, \text{this a pen})\}$$
$$(\epsilon, \text{is a pen})/\Sigma=(\epsilon, \text{is a desk})/\Sigma=(\epsilon, \text{this a pen})/\Sigma$$
$$(\text{is, a pen})/\Sigma=(\text{this, a pen})/\Sigma=(\text{that, a desk})/\Sigma \neq (\epsilon, \text{is a desk})/\Sigma$$
$$C_{\text{that}}{}^{\Sigma}(\text{this})=\{(\epsilon, \text{is a pen})\}$$
$$C_{\text{this}}{}^{\Sigma}(\text{that})=\{(\epsilon, \text{is a desk})\}$$

すなわち，

$$n(C(\text{this}))=2$$
$$n(C(\text{that}))=1$$
$$n(C_{\text{that}}{}^{\Sigma}(\text{this}))=1$$
$$n(C_{\text{this}}{}^{\Sigma}(\text{that})=1$$

ゆえに，

$$F_{\Sigma}(\text{this, that})=\frac{1+1}{2+1}=\frac{2}{3}$$

また，

$$C_{\text{that}}{}^{\Sigma}(\text{is})=\{(\epsilon, \text{this a pen})\}$$
$$C_{\text{is}}{}^{\Sigma}(\text{that})=\{\epsilon, \text{is a desk})\}$$

すなわち，

$$n(C(\text{is}))=3$$
$$n(C_{\text{that}}{}^{\Sigma}(\text{is}))=1$$
$$n(C_{\text{is}}{}^{\Sigma}(\text{that}))=1$$

ゆえに，

$$F_{\Sigma}(\text{that, is})=\frac{1+1}{3+1}=\frac{2}{4}$$

ここで，

$R=\{\text{this, that}\} \cup \{\text{is}\} \cup \{\text{a}\} \cup \{\text{pen, desk}\}$

とおけば，

$$(\epsilon, \text{is a desk})/R \neq (\epsilon, \text{this a pen})/R$$

よって，

$$C_{\text{that}}{}^{R}(\text{is})=C_{\text{is}}{}^{R}(\text{that})=\phi$$

ゆえに，

$$F_R(\text{that, is})=\frac{0+0}{3+1}=0$$

ただし，ϵ は空文を，分割 R の { } は1つの細胞（同値類）を表わしている．

〔**定義 2.4**〕 一致の度合 $F_R(a, b)$ が $\lambda(0 \leqq \lambda \leqq 1)$ 以上のとき，2つの単語 a と b は関係 $A_\lambda(R)$ があるという．$A_\lambda(R)$ の反射的推移的閉包 $A_\lambda{}^*(R)$ を R から λ 誘導された分割といい，分割 R から λ 誘導を n 回繰り返して得られる分割を ${}_nA_\lambda{}^*(R)$ と記す．

関係 $A_\lambda(R)$ は必ずしも同値関係ではないが対称的であるので，$A_\lambda{}^*(R)$ は同値関係（分割）である．また，定理 2.5 からわかるように，$F_R(a, b)=1$ と $a \cdot \sim/R \cdot b$ は同値である．したがって，関係 $A_{\lambda=1}(R)$ は同値関係であり，$A_{\lambda=1}(R)=\sim/R$

〔**定義 2.5**〕 分割 R と R から λ 誘導された分割 $A_\lambda{}^*(R)$ が等しいとき，分割 R は λ 適合しているという．

〔**定義 2.6**〕 分割 R から λ 誘導を繰り返したとき同一の分割しか得られなくなった（λ 適合した）とき，その λ 誘導分割列（または分割 R）は λ 収束するといい，その分割を ${}_\infty A_\lambda{}^*(R)$ と表わす．

文献 1）における誘導と適合に関する定理と同様な定理が λ 誘導と λ 適合に関しても成立する．

〔**定理 2.6**〕 $R \geq A_\lambda{}^*(R)$（または，$R \leq A_\lambda{}^*(R)$）

ならば R は λ 収束する.

とくに，非本来の分割 Σ と単位分割 E は任意の分割 R に対して $\Sigma \geq R \geq E$. ゆえに，Σ および E は任意の $\lambda(0<\lambda\leq 1)$ に対して λ 収束する．ただし，λ 誘導の回数は1回とは限らない.

収束する誘導分割列に対して以下の定理が成立する.

〔定理 2.7〕 $R \leq R'$ ならば

$$A_\lambda^*(R) \leq A_\lambda^*(R')$$
$$_nA_\lambda^*(R) \leq {}_nA_\lambda^*(R') \quad (n=1,2,\cdots)$$

ゆえに,

$$_\infty A_\lambda^*(R) \leq {}_\infty A_\lambda^*(R')$$

〔定理 2.8〕 $\lambda \leq \lambda'$ ならば

$$A_\lambda(R) \geq A_{\lambda'}(R)$$
$$A_\lambda^*(R) \geq A_{\lambda'}^*(R)$$
$$_\infty A_\lambda^*(R) \geq {}_\infty A_{\lambda'}^*(R)$$

〔定理 2.9〕

$$R \geq R' \geq {}_\infty A_\lambda^*(R) \;\text{または}\; ({}_\infty A_\lambda^*(R) \geq R' \geq R)$$

ならば R' は λ 収束して

$$_\infty A_\lambda^*(R) = {}_\infty A_\lambda^*(R')$$

〔証明〕 $R \geq R' \geq {}_\infty A_\lambda^*(R)$ の場合を証明する．定理 2.7 を用いて証明する.

$R \geq R'$. ゆえに，${}_nA_\lambda^*(R) \geq {}_nA_\lambda^*(R')$.

$R' \geq {}_\infty A_\lambda^*(R)$. ゆえに，${}_\infty A_\lambda^*(R)$ が λ 適合していることを考慮に入れれば,

$$_nA_\lambda^*(R) \geq {}_nA_\lambda^*({}_\infty A_\lambda^*(R)) = {}_\infty A_\lambda^*(R)$$

よって，R が N 回の λ 誘導で収束する (${}_NA_\lambda^*(R) = {}_\infty A_\lambda^*(R)$) とすれば,

$$_\infty A_\lambda^*(R) = {}_NA_\lambda^*(R) \geq {}_NA_\lambda^*(R') \geq {}_\infty A_\lambda^*(R)$$

この式は R' から N 回 λ 誘導した分割 ${}_NA_\lambda^*(R')$ は ${}_\infty A_\lambda^*(R)$ に等しく，λ 適合していることを示している．ゆえに，${}_\infty A_\lambda^*(R) = {}_\infty A_\lambda^*(R')$.

(証明終)

〔定理 2.10〕 $\lambda \leq \lambda'$ のとき

$\Sigma \geq {}_\infty A_\lambda^*(\Sigma) \geq {}_\infty A_{\lambda'}^*(\Sigma)$ より

$$_\infty A_{\lambda'}^*({}_\infty A_\lambda^*(\Sigma)) = {}_\infty A_{\lambda'}^*(\Sigma)$$

〔定理 2.11〕 $\lambda \leq \lambda'$ のとき

$E \leq {}_\infty A_{\lambda'}^*(E) \leq {}_\infty A_\lambda^*(E)$ より

$$_\infty A_\lambda^*({}_\infty A_{\lambda'}^*(E)) = {}_\infty A_\lambda^*(E)$$

以上の定理より，たとえば，非本来の分割 Σ から出発して，$\lambda=0.2$, $\lambda=0.4$ に関して ${}_\infty A_\lambda^*(\Sigma)$ を求める場合には，まず，${}_\infty A_{\lambda=0.2}^*(\Sigma)$ を求め，$\lambda=0.4$ の場合の収束する分割 ${}_\infty A_{\lambda=0.4}^*(\Sigma)$ を求めるときには，Σ から λ 誘導 ($\lambda=0.4$) を繰り返すかわりに ${}_\infty A_{\lambda=0.2}^*(\Sigma)$ から λ 誘導 ($\lambda=0.4$) を繰り返して ${}_\infty A_{\lambda=0.4}^*(\Sigma) = {}_\infty A_{\lambda=0.4}^*({}_\infty A_{\lambda=0.2}^*(\Sigma))$ を求めてもよいことがわかる．また，その方が繰り返す λ 誘導の回数が少ない可能性がある（定理 2.9).

3. λ 適合分割を求めるアルゴリズム

この章では与えられた分割から λ 誘導を繰り返して λ 収束させ λ 適合する分割を求めるアルゴリズムの概略を説明する.

Fig. 1 に処理全体の流れ図を示す.

手順 1. 言語（文の集まり）を定める.

何を単語，文，言語としてとらえるか定め，文の集まりを計算機に与え，単語を切り出す方法を与える．単語の表および文の表を作る.

この手順1では処理する目的によって，単語，文，

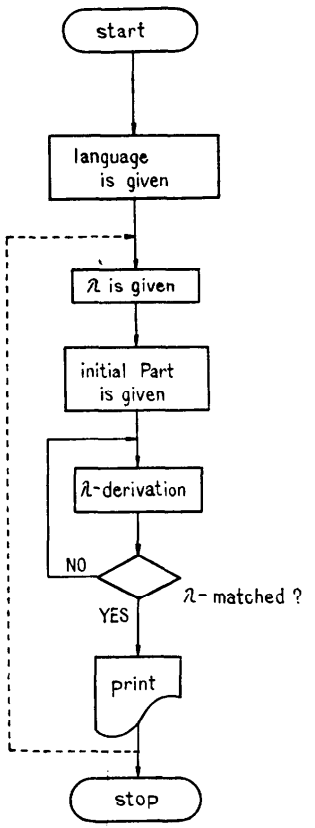

Fig. 1 General flowchart of seaching for λ-matched partition

言語の定義を設定しなければならない。たとえば，単語では大文字と小文字の区別，複数形，省略形などをどのように扱うかを決定しなければならない。文では大文字で始まり終止符（.）などで終るものを文と考える方法，段落を文と考える方法，一冊の本を文と考える方法などがある．言語では共通語だけを考える方法，方言をも含めて考える方法などがある．

手順 2. λ および初期分割を定める．

初期分割の設定には，どのような目的で処理を行うか（処理水準）を考慮して，単位分割，語源が同じ語を同一細胞に入れる分割，単複などの変形があっても同一細胞に入れる分割，非本来の分割などを選ばなければならない．また，λ の設定には一致の度合はどの程度でよいかを考えて設定しなければならないが，ある初期分割において，ある範囲の λ に対してかなり安定した結果が得られると推察できる場合にはその範囲の λ を設定した方がよいと思われる．

ここでは分割を表現する際には，単語とその単語が属する細胞の番号を組にして表わす．λ 誘導を繰り返すので **Fig. 2(a)** の表を用意する．ただし，λ 収束することがわかっている場合には **Fig. 2(b)** の表を用いることができる．

ここで，初期分割が非本来の分割 Σ（単位分割 E）のときには，λ を小さい（大きい）方から順に与えれば，2回目からは λ の小さな（大きな）値で λ 収束した分割を初期分割にしても定理 2.10（定理 2.11）より得られる結果は同じであり，λ 収束に要する λ 誘導の回数が少なくなるから小さな（大きな）値の λ において λ 収束した結果を大きな（小さな）値の λ に対する初期分割とすればよい．

手順 3. **Fig. 3** に手順3の少し詳しい流れ図を示す．

すべての2つの単語 a_i, a_j の組に対して

$$F_R(a_i, a_j)=\frac{n(C^R{}_{a_j}(a_i))+n(C^R{}_{a_i}(a_j))}{n(C(a_i))+n(C(a_j))}$$

を計算する．ただし，分割が λ 誘導を繰り返すたびに

word	derivative No.		
	1	2	3
this	1		
that	1		
is	2		
a	3		

(a)

word	old cell No	new cell No

(b)

Fig. 2 Word-partition table

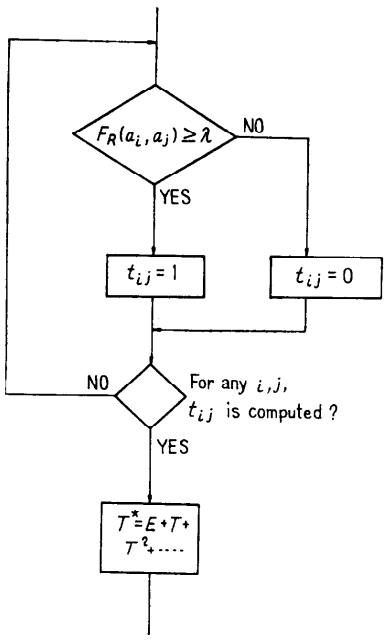

Fig. 3 Flowchart of λ-derivation

小さくなることがわかっている場合にはすべての a_i, a_j の組について $F_R(a_i, a_j)$ を求める必要はなく，a_i, a_j が R 同値な場合のみに求めればよい．

$F_R(a_i, a_j)\geq\lambda$ ならば $t_{ij}=1$, $F_R(a_i, a_j)<\lambda$ ならば $t_{ij}=0$ として行列 $T=(t_{ij})$ を作る．この T は $A_\lambda(R)$ を表わしているので $A_\lambda{}^*(R)$ を求めるには T のブール積のブール和 $T^*=E+T+T^2+\cdots$ を求めればよい．T^* の (i, j) 要素を $t_{ij}{}^*$ とすれば，$t_{ij}{}^*=1$ と $a_i \cdot A_\lambda{}^*(R) \cdot a_j$ は同値である．この $A_\lambda{}^*(R)$ を単語表に細胞の番号を入れることで記入する．

手順 4. λ 収束したかどうかを確かめる．

λ 誘導を繰り返したとき λ 収束しない（最後に同期的になる）場合もあるので λ 収束の判定と同時に最後に周期的になっているかどうかの判定もしなければならない．λ 誘導についても文献 1) の定理 3.7 と同様の定理が成立するので，λ 収束しない場合は最後の周期の部分の和の反射的推移的閉包および積を初期分割として λ 誘導を繰り返せば λ 収束し，λ 適合した分割が得られる．λ 誘導分割列が λ 収束することがわかっているときには，R と $A_\lambda{}^*(R)$ が一致しているかどうかを確かめればよい．λ 収束していない周期的にもなっていない場合には $A_\lambda{}^*(R)$ を R として(Fig, 2 (b) の表を用いる場合は New partition を Old partition に移して）手順3に戻る．

手順 5. λ 収束すれば λ の値，初期分割，λ 収束した分割を書き出す．

4. 実験結果

この章では前章のアルゴリズムに基づいて英語の教科書を対象として行った実験の結果について述べる．

手順1における言語，文，単語はつぎのように与えている．

言語は中学1年の英語の教科書 (New Prince Readers, 開隆堂，昭 43) の Lesson 1 から Lesson 3 までの本文である．

文の大文字で始まり，終止符(.)または疑問符(?)で終る．

単語はつぎの手順で切り出す．

(1) 大文字，小文字の区別はしない．

(2) 省略形はもとの形になおす．

たとえば，

isn't→is not

that's→that is

(3) 空白ではさまれたアルファベットの系列を単語とみなす．ただし，コンマ(,)，終止符，疑問符も1つの単語とみなす．

手順2の初期分割を非本来の分割にして，λ=0.2, 0.4, 0.6 に対して収束した分割を **Table 1** に示す．この表のわくは各細胞の区切りを表わしている．

得られた結果について検討する．

(1) λ=0.4 の場合と λ=0.6 の場合結果(収束した分割) は等しく，λ の変化に対してかなり安定である．

Table 1 Examples of λ-matched partitien (maximum member)

λ=0.4　λ=0.6	λ=0.2
.,?	., ?
what	what
too	too
or	or
good, old, small, big, bad	good old, small big, bad
not	not
, (comma)	,
no	no
yes	yes
a an	a, an
is, it, this, that	is, it, this, that, apple, orange, pen, desk, hat, ball, cat, lemon, cap, bag, bat, mitt, glove, dog, sheep, bed, pencil, basket table, peach, tomato.
apple orange pen, desk, hat, ball, cat, lemon, cap, bag, bat, mitt, glove, dog, sheep, bed pencil, basket, table, peach, tomato.	

(2) λ=0.4, 0.6 に対して収束した分割は品詞分割と考えてよいほど期待した結果に近い．

(3) しかし，is と this, that, it が同一の細胞に属している．これは肯定文と疑問文が単に is と this などの語順を変化させ，終止符と疑問符を置き換えただけであるためである．

このように文の類に対して一定の語順操作を行ってもその言語の文になる場合は異なった細胞に入れるべき2つの単語が最大適合分割では同一の細胞に入る場合がある．

この欠点を取り除くためにはつぎの2つが考えられる．

(1) 文を基本的な構造を持つだけに限り，変形された文を除外する．

(2) 論理的な単語は1つの単語で1つの細胞を作ると仮定する．

以上のことを考え，実験に用いた文の中で平叙文だけを取り出して同様の実験を行った結果が **Table 2** である．

Table 1 と Table 2 を比較すれば，平叙文だけで実験した場合が非常によい結果が出ていることがわかる．

つぎに，単位分割 E を初期分割として λ 誘導を繰り返して得られた λ 適合する分割を表わしているのが **Table 3** (次頁参照) である．

この表よりわかるように冠詞 a と an が異なった細胞に入っている．これは a の後に来る名詞は子音で始まり，an の後に来る名詞は母音で始まっていることに帰因する．このように単位分割を初期分割として得られた適合する分割は同一の品詞列でも発音などの影響を受けて a と an が異なる細胞に入るなど直接

Table 2 An example of λ-matched partition (maximum member)
λ=0.2, 0.4, 0.6

.
too
a, an
not
small, big, bad, old
,
is
it, this, that
apple, orange, pen, desk, hat, ball, cat, lemon, cap, bag, bat, mitt, glove, dog, sheep, bed, pencil, basket, table, peach, tomato.

Table 3 An example of λ-matched partition (minimum member)
λ=0.1

.
too
a
an
not
old
small, big, bad
,
is
it, this, that
apple, orange
pen, desk, hat, ball, cat, lemon, cap, bag, bat, mitt, glove, dog, sheep, bed, pencil, basket, table, peach, tomato.

シンタックス（品詞）と無関係な要因が混入している恐れがある．

5. むすび

文脈の R の意味での一致を定義し，λ 適合する分割を求めるアルゴリズムを示し，簡単な実例を示した．しかしながら，与えられた初期分割から λ 誘導を繰り返したとき λ 収束するかどうかの判定条件を求めることなどは今後に残された問題である．また，変形の問題，意味の問題など残された問題は多い．これらについては他の機会に報告したい．

最後に，近畿大学・理工学部・滝猪一教授の御援助に感謝する．

参考文献

1) 佐藤睦，田中幸吉：語いの一分割法の提案，情報処理，Vol. 16, No. 2, pp. 102～107 (1975).
2) S. Marcus: Algebraic Linguistics; Analytical Models, Academic Press, New York & London (1967).
3) F. Kiefer: Mathematical Linguistics in Eastern Europe, Elsevier, New York (1968).

(昭和50年7月14日受付)
(昭和50年11月29日再受付)